Eichner, Thomas; Pethig, Rüdger

Working Paper

Self-enforcing environmental agreements and international trade

CESifo Working Paper, No. 4125

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Eichner, Thomas; Pethig, Rüdger (2013) : Self-enforcing environmental agreements and international trade, CESifo Working Paper, No. 4125, Center for Economic Studies and ifo Institute (CESifo), Munich

This Version is available at:
https://hdl.handle.net/10419/69987

The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

Working Papers
www.cesifo.org/wp

Self-Enforcing Environmental Agreements and International Trade

Thomas Eichner
Rüdiger Pethig

CESIFO WORKING PAPER NO. 4125
CATEGORY 10: ENERGY AND CLIMATE ECONOMICS
FEBRUARY 2013

Presented at CESifo Area Conference on Energy & Climate Economics, November 2012

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 4125*

Self-Enforcing Environmental Agreements and International Trade

Abstract

In the basic model of the literature on international environmental agreements (IEAs) (Barrett 1994; Rubio and Ulph 2006) the number of signatories of self-enforcing IEAs does not exceed three, if non-positive emissions are ruled out. We extend that model by introducing a composite consumer good and fossil fuel that are produced and consumed in each country and traded on world markets. When signatory countries act as Stackelberg leader and emissions are positive, the size of stable IEAs may be significantly larger in our model with international trade. This would be good news if larger self-enforcing IEAs would lead to stronger reductions of total emissions. Unfortunately, in the presence of self-enforcing IEAs total emissions turn out to be only slightly less than in the business as usual scenario, independent of the number of signatories. We also investigate the role of international trade by comparing our free-trade results with the outcome in the regime of autarky. Our autarky model turns out to coincide with the basic model of the literature alluded to above. We contribute to that literature by showing that in autarky the outcome of self-enforcing IEAs is also almost the same as in business as usual.

JEL-Code: C720, F020, Q500, Q580.

Keywords: international trade, self-enforcing environmental agreements, Stackelberg equilibrium.

Thomas Eichner
Department of Economics
University of Hagen
Universitätsstraße 41
Germany – 58097 Hagen
thomas.eichner@fernuni-hagen.de

Rüdiger Pethig
Department of Economics
University of Siegen
Hölderlinstraße 3
Germany – 57068 Siegen
pethig@vwl.wiwi.uni-siegen.de

CESifo sponsorship for presenting the paper at the CESifo Area Conference on Energy and Climate Economics, November 2012, and helpful comments from an associate editor and three anonymous referees are gratefully acknowledged. Remaining errors are the authors' sole responsibility.

1 The problem

International environmental agreements (IEAs) are essential for the stabilization of the world climate at safe levels through the effective reduction of global carbon emissions. The first legally binding international agreement on climate protection, the Kyoto Protocol, has been criticized because it includes commitments for a small number of countries only and is therefore likely to accomplish very little in terms of global emission reduction (Buchner et al. 2002). The prospects are bleak for reaching an IEA in the near future which accomplishes both attracting many signatories and reducing global emissions significantly. The tedious practical negotiations and the serious global change challenge call for continued investigations of the theoretical foundations of successful and effective IEAs.

Since the early 1990s an economic literature has developed on self-enforcing IEAs. An IEA is said to be self-enforcing or stable if no signatory country has an incentive to leave the IEA and no non-signatory country has an incentive to join. The seminal papers on self-enforcing IEAs include Barrett (1992, 1994), Hoel (1992) and Carraro and Siniscalco (1993). Most papers are quite pessimistic about the stability of large IEAs. Carraro and Siniscalco (1991), Hoel (1992) and Finus (2001) find that a stable IEA consists of three countries when the climate damage is linear and of two countries when the climate damage is quadratic. These papers assume that both signatories and non-signatories behave in a Cournot-Nash fashion.

Another strand of the IEA literature which we will follow in the present paper makes use of the Stackelberg assumption portraying the climate coalition[1] as Stackelberg leader and all non-cooperative countries as Stackelberg followers. In that framework Barrett's (1994) simulation results suggest the existence of stable coalition sizes between two and the grand coalition. However, Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006) proved that large stable IEAs imply zero emissions (corner solutions) or negative emissions. The latter must clearly be ruled out in models without stock pollution, because it is infeasible to abate more emissions than are generated. As Rubio and Ulph point out, the reason for negative emissions is the assumption of non-essential emissions which is standard in the literature on IEAs. Although that assumption is unrealistic in the case of carbon emissions and climate change, we will stick to it for reasons of tractability and comparability with pertaining literature and we will follow Diamantoudi and Sartzetakis (2006) in restricting parameter values such that the resultant emissions are always strictly positive. Under that constraint (along with the assumption of non-essential emissions)

[1] In the present paper we use the terms IEA and (climate) coalition as synonyms. Our exclusive focus is on a single coalition.

Diamantoudi and Sartzetakis (2006) as well as Rubio and Ulph (2006) find that the number of signatories of self-enforcing IEAs is not larger than four.

The *basic model* of an IEA employed by Barrett (1994), Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006) and others is a static model of symmetric countries where each country's domestic emissions generate domestic welfare that is decreasing at the margin and where all countries' emissions create a welfare loss (climate damage) which is uniform across countries and increasing at the margin.[2] That model does not account for production, consumption, markets and international trade and thus captures the world economy in a rudimentary way only. It has been extended in various directions (Finus 2003).[3] For example, Hoel and Schneider (1997) introduce transfer schemes in the coalition formation process, Kolstad (2007) studies systematic uncertainty and Carbone et al. (2009) use the *basic model* for an empirical investigation of how international emission trading impacts on IEAs. However, we are not aware of studies on the formation of IEAs[4] that model in more detail the economies of individual countries and their economic interdependencies.

The present paper aims to extend the *basic model* along these lines and then investigates the impact of that extension on the stability of IEAs in the Stackelberg leader-follower framework. We will add structure to the national economies by introducing a consumer good and fossil fuel that are produced in each country, consumed by its representative consumer and traded on world markets.[5] In this general equilibrium framework we first briefly characterize the business-as-usual (BAU) scenario with non-cooperative governments as a benchmarkt and then turn to our central theme, the characterization of self-enforcing IEAs in the Stackelberg model.

For the case of positive equilibrium emissions we find that - depending on parameter constellations - international trade may significantly increase the size of stable IEAs. That is, the conditions for successful sub-global cooperative action appear to be more favorable than suggested by the *basic model* of the IEA literature. Unfortunately, the hope for a more

[2]Barrett (1994) models abatement, and therefore his approach seems to differ from the *basic model*, at first glance. However, as pointed out by Diamantoudi and Sartzetakis (2006, Section 4), Barrett's model is equivalent to the *basic model* as long as abatement does not exceed the flow of emissions.

[3]Modifying and extending, respectively, the *basic model*, Barrett (1999) and Hannesson (2010) show that stable coalitions may consist of a large number of countries, if the coalition behaves as a Nash player.

[4]There are also studies relaxing the assumption of the basic model that countries are identical (e.g. Barrett 2001). In the present paper we will stick to that assumption to keep our model tractable.

[5]Despite the importance of international trade for the formation of IEAs, to our knowledge there is only one paper dealing with that issue, and that is Barrett (1997) who illustrates in a partial equilibrium model with abatement how trade policy may help support stable IEAs. Copeland and Taylor (2005) study the role of international trade in a model of non-cooperative heterogenous countries coping with a global (climate) externality. They do not address the formation of coalitions, however.

optimistic view on *effective* cooperative emission reductions turns out to be unwarranted because our second main finding is that if an IEA of any size is self-enforcing, the corresponding level of world emissions is only slightly lower than in business as usual (BAU). These results are obtained in a very simple model making use of parametric functions and numerical examples and they may not be, therefore, reliable indicators for the outcome of the highly complex ongoing international climate negotiations. Nonetheless, they provide some support for the disturbing view that attempts to form a sub-global climate coalition (of whatever size) are futile.

As the introduction of international trade represents a major extension of the IEA literature, it is natural to highlight its impact on results by investigating the outcome of our model in the absence of international trade (autarky). When all countries are autarkic, our model turns out to coincide with the *basic model* of the literature on IEAs which has established, as reported above, that the number of signatories in self-enforcing IEAs is very small. We find that turning from free trade to autarky reduces the size of stable coalitions for any given set of parameters. Moreover, we extend the literature by showing that in the autarky scenario - as under free trade - the level of world emissions is only slightly lower than in BAU.

The paper is organized as follows. Section 2 introduces the model and briefly analyzes the business-as-usual scenario which serves as a benchmark throughout the paper. Section 3.1 prepares for the analysis of self-enforcing IEAs in Section 3.2 by characterizing the outcome of the Stackelberg game and its dependence on the size of coalitions. Section 4 deals with the role of international trade for the results by comparing the regimes of free trade and autarky and by linking the case of autarky to the basic model of the coalition formation literature. Section 5 concludes.

2 The model

The world economy consists of n identical countries. Each country produces two consumer goods. The first is a standard composite good, called *good X* (quantity x_i) and the second is a fossil energy carrier (quantity e_i), e.g. oil, gas or coal extracted from domestic fossil reserves. We refer to that good simply as *fuel*.[6] Each country's production technology is

[6]Households do not consume fuel directly but use fuel as input in a linear household production function to produce e.g. the commodities heat or transportation services. To keep the exposition simple, we refrain from modeling the household production technology, however, and interpret fuel as consumer good.

represented by the production possibility frontier[7]

$$x_i^s = T(e_i^s) \quad i = 1, \ldots, n, \tag{1}$$

where the function T is decreasing and strictly concave in e_i^s. The transformation function (1) implies that both commodities are produced by means of domestic productive factors (e.g. labor and capital) whose endowments are given. The utility[8]

$$V(e_i^d) + x_i^d - D\left(\sum_j e_j^d\right) \tag{2}$$

of the representative consumer of country i is additive separable in all arguments and linear in the consumption x_i^d of good X. V is increasing and concave, and D is increasing and convex in its argument. The consumption of fuel generates the greenhouse gas carbon dioxide whose emission is proportional to fuel consumption. Emission units are chosen such that e_i^d denotes both fuel demanded by consumer i and carbon emissions from burning fuel. There is no abatement technology for emission reduction.[9] The function D captures the climate damage caused by worldwide carbon emissions from burning fuel.

For the sake of more specific results, throughout the paper we will specify the functions T, V and D from (1) and (2) by the following quadratic functional forms:[10]

$$T(e_i^s) = \bar{x} - \frac{\alpha}{2}(e_i^s)^2, \quad V(e_i^d) = ae_i^d - \frac{b}{2}(e_i^d)^2, \quad D\left(\sum_j e_j^d\right) = \frac{1}{2}\left(\sum_j e_j^d\right)^2, \tag{3}$$

where $\bar{x}$, a, b and α are positive parameters.

In our stylized model (1) and (2) of the individual country's economy all fuel goes from production directly to consumers where 'fuel production' can be interpreted to include extraction of fossil energy carriers as well as production of electricity, gasoline, gas or coal for non-business usage.[11] Although in practice climate regulation does not only apply to the consumers' energy demand but also to energy-consuming industries, as e.g. in the

[7]The superscript s indicates quantities supplied. Upper-case letters denote functions. Subscripts attached to them indicate partial derivatives.

[8]The superscript d indicates quantities demanded.

[9]Carbon capture and sequestration is a potential abatement technology which is unlikely to be applied on a large scale in the near or medium term future.

[10]In (3) the parametric form of $T(e_i^s)$ can be 'microfounded' as follows. Let $\bar{r}$ be country i's endowment of a (composite) production factor and consider the production functions $x = \alpha_x r_x$ and $e = (r_e/\alpha_e)^{1/2}$ with $r_e + r_x = \bar{r}$. α_e, α_x are positive constants. The quadratic transformation function in (3) is straightforward from these three equations when setting $\bar{x} := \alpha_x \bar{r}$ and $\alpha := \alpha_x \alpha_e$.

[11]Such simplifications are driven by limits of tractability. We also wish to recall, however, that the model of the present paper is far more complex than the *basic model* of IEA (e.g. Finus 2003, Section 2.3) which does without specifying production, consumption and markets, as we have pointed out in the introduction.

EU emission trading scheme, we maintain that our simplification still captures the central issue of emission regulation. Whether fuel consumption of industries or of consumers is regulated, in both cases more stringent emission caps require raising the domestic price for fuel consumption which, in turn, induces allocative displacement effects via changes in relative prices.

There are perfectly competitive world markets for good X (price $p_x \equiv 1$) and for fuel (producer price p), and the markets are in equilibrium if

$$\sum_j x_j^s = \sum_j x_j^d \quad \text{and} \quad \sum_j e_j^s = \sum_j e_j^d. \tag{4}$$

The firms' supply of fuel is straightforward. Taking prices as given, the (aggregate) producer i maximizes profits $x_i^s + pe_i^s$ subject to (1) which yields the first-order condition

$$p = -T'(e_i^s) \quad \text{for} \quad i = 1, \ldots, n. \tag{5}$$

Combined with (1), equation (5) implies a fuel supply function

$$e_i^s = E^s(p) \quad \text{with} \quad E_p^s > 0 \quad \text{for} \quad i = 1, \ldots, n. \tag{6}$$

Each government i regulates domestic carbon emissions by enforcing an emission cap e_i. For the time being we suppose these caps are arbitrarily fixed and tight enough to be binding. To implement its emission cap, government i issues the amount e_i of emission permits and auctions them at the permit price π_i. Consumers in country i need to acquire emission permits to match their purchase of fuel. The representative consumer i ignores the impact of her emissions on climate damage and maximizes her (consumption) utility $V(e_i^d) + x_i^d$ subject to her budget constraint

$$x_i^d + (p + \pi_i)e_i^d = y_i, \quad \text{where} \quad y_i := x_i^s + pe_i^s + \pi_i e_i^d \tag{7}$$

is consumer i's income (= profit income plus recycled revenues from the permit auction). From the first-order condition $p + \pi_i = V'(e_i^d)$ follows a fuel demand function

$$e_i^d = E^d(p + \pi_i) \quad \text{for} \quad i = 1, \ldots, n. \tag{8}$$

The result of auctioning the permits obviously is

$$e_i^d = e_i \quad \text{for} \quad i = 1, \ldots, n. \tag{9}$$

Combining the equilibrium condition $\sum_j e_j^s = \sum_j e_j^d$ from (4) with (6) and (9) yields

$$e_i^s = E^s(p) = \frac{\sum_j e_j}{n} \quad \text{for} \quad i = 1, \ldots, n. \tag{10}$$

Equation (10) determines the unique equilibrium price of fuel and also establishes that in equilibrium all firms produce the same amount of fuel, $\sum_j e_j/n$. From (5), (8) and (9) follows $e_i = E^d\left[-T'\left(\frac{\sum_j e_j}{n}\right) + \pi_i\right]$. This equation determines the unique equilibrium permit price. The equilibrium supplies and demands on the market for good X are

$$x_i^s = T\left(\frac{\sum_j e_j}{n}\right) \quad \text{and} \quad x_i^d = T\left(\frac{\sum_j e_j}{n}\right) - T'\left(\frac{\sum_j e_j}{n}\right)\left(\frac{\sum_j e_j}{n} - e_i\right), \tag{11}$$

where the first equation in (11) is implied by (1) and (10) and the second by (1), (7), (9) and (10). It readily follows from (11) that the market for good X is in equilibrium, if the fuel market is in equilibrium.

To sum up, in the world economy with non-cooperative emission cap regulation there is a unique competitive equilibrium for every profile $(e_1, \ldots, e_n)$ of binding emission caps. That is, in equilibrium all demands and supplies, and all prices are determined by $(e_1, \ldots, e_n)$. Combining welfare (2) with (9), (10) and (11) results in the equilibrium welfare of country $i = 1, \ldots, n$,

$$W^i(e_1, \ldots, e_n) := V(e_i) + T\left(\frac{\sum_j e_j}{n}\right) - \left(\frac{\sum_j e_j}{n} - e_i\right)T'\left(\frac{\sum_j e_j}{n}\right) - D\left(\sum_j e_j\right). \tag{12}$$

So far we have considered governments that fix national emission caps in an arbitrary way. From now on their objective function is supposed to be national welfare, (12). Before addressing cooperation in emission regulation, we briefly investigate the benchmark case of global non-cooperation. In game-theoretic language, the n governments are the players of a non-cooperative game. Their strategies are national emission caps and their payoff functions are national welfares, (12). The natural solution concept is the Nash equilibrium, a state, where each government's emission cap is the best response to each other government's emission cap. We refer to that equilibrium as business as usual (BAU). In terms of the formal model, government i chooses that cap e_i which maximizes $W^i(e_1, \ldots, e_n)$ for given caps $(e_1, \ldots, e_{i-1}, e_{i+1}, \ldots, e_n)$. Differentiation of (12) with respect to e_i yields the first-order condition[12] $W^i_{e_i} = 0$.

Eichner and Pethig (2012) show that $W^i_{e_i} = 0$ can be converted into a best reply function $\tilde{R}$ satisfying

$$e_i = \tilde{R}\left(\sum_{j\neq i} e_j\right) \tag{13}$$

whose first derivative is in the interval $]-1, 0[$ under mild restrictions. Hence there exists a unique symmetric Nash equilibrium satisfying $e_i = e_j$ for all $j \neq i$. If the functions

[12]Throughout the paper we restrict our attention to interior solutions.

V, T and D are specified as in (3) the uniform Nash equilibrium cap is $e_o := \frac{a}{\alpha+b+n}$. Another immediate consequence of symmetry is that international trade does not take place in equilibrium. Each country sets its BAU emission cap e_o such that its marginal benefit of consumption, $V'(e_o) + T'(e_o)$, equals its marginal climate damage, $D'(ne_o)$. If the countries would disregard their own impact on climate damage, national equilibrium emissions would exceed e_o. Hence in BAU some emission reduction is in the countries' self-interest. It is also clear that total emissions ne_o in BAU exceed total emissions in the optimal fully cooperative solution, since all countries disregard in BAU the positive external effects of their emission reduction on the other countries.

3 Climate coalition as Stackelberg leader

Suppose now that some countries are members in a climate coalition, whereas all other countries continue to act non-cooperatively. For the sake of formal analysis, we lump together the first m countries, $2 \leq m < n$, in one group, denoted group $C := \{1, 2, \ldots, m\}$ with C for coalition, and collect all remaining countries in another group, denoted group $F := \{m+1, \ldots, n\}$ with F for fringe. Our focus will be on a game of sequential choice of emission caps in which the coalition is the Stackelberg leader and moves first and the fringe countries are Stackelberg followers. The coalition formation literature has made ample use of the Stackelberg assumption (Finus 2001) and we refer the reader to that literature for information on the discussion about the plausibility and relative merits of the Nash concept on the one hand and the Stackelberg concept on the other.[13] Our aim is to investigate how the Stackelberg assumption drives the outcome of the game when we extend the basic model as outlined in Section 2.

3.1 Climate coalitions and coalition sizes

Stackelberg equilibrium In the present section we aim to characterize the allocation in the Stackelberg equilibrium (to be specified below) for alternatively given coalition sizes and thus prepare for the analysis of coalition stability in the next Section 3.2. The objective of the climate coalition C is to maximize the joint welfare $\sum_{j \in C} W^j(e_1, \ldots, e_n)$ of its members taking the behavior of the fringe countries into account. Since all coalition countries are alike, $e_i = e_j$ for all $i, j \in C$ is a necessary maximum condition which allows us to set $e_i = e_c$ for all $i \in C$. Thus the coalition can be treated as a single player whose strategy will be denoted

[13] Eichner and Pethig (2012) is a companion paper in which the climate coalition is modeled as a Nash player.

as $s_c := me_c$. We continue portraying fringe countries as non-cooperative Nash players, and therefore $W^i_{e_i} = 0$ still applies for each fringe country. As $W^i_{e_i} = 0$ cannot be satisfied for $i, j \in F, i \neq j$, unless $e_i = e_j$, we proceed by setting $e_i = e_f$ for all $i \in F$. With this notation, each fringe country's best-reply function (13) reads $e_f = \tilde{R}[s_c + (n-m-1)e_f]$. Eichner and Pethig (2012) show that this equation implies a function R satisfying $(n-m)e_f = R(me_c, m)$ or

$$s_f = R(s_c, m) \quad \text{with} \quad R_{s_c} \in]-1, 0[, \tag{14}$$

where $s_c := me_c$ and $s_f := (n-m)e_f$.

According to (14) fringe countries can be treated as if they act as a single player whose strategy is s_f. In that sense R is the 'aggregate' best reply function of 'the fringe'. However, it is important to emphasize that R is a purely formal transformation of $\tilde{R}$ from (13), and therefore (14) does not imply any cooperation among fringe countries. R turns out to be an important analytical tool.

With the newly introduced notation $s_f := (n-m)e_f$, we next express total emissions as $\sum e_j = s_c + s_f$ and rewrite the welfare of individual countries, (12), as

$$W^c(s_c, s_f, m) := V\left(\frac{s_c}{m}\right) + T\left(\frac{s_c + s_f}{n}\right) - \left(\frac{s_c + s_f}{n} - \frac{s_c}{m}\right) T'\left(\frac{s_c + s_f}{n}\right) - D(s_c + s_f) \tag{15}$$

for all countries in group C and as

$$\begin{aligned} W^f(s_c, s_f, m) \;:=\; & V\left(\frac{s_f}{n-m}\right) + T\left(\frac{s_c + s_f}{n}\right) - \left(\frac{s_c + s_f}{n} - \frac{s_f}{n-m}\right) T'\left(\frac{s_c + s_f}{n}\right) \\ & - D(s_c + s_f) \end{aligned} \tag{16}$$

for all countries in group F. For convenience of notation and later reference we refer to $(-D(s_c + s_f))$ as the *climate welfare* and to $W^j(s_c, s_f, m) + D(s_c + s_f)$ as the *consumption welfare* of an individual country.

Being the Stackelberg leader the coalition of size $m \in \{1, \ldots, n\}$ accounts for (14) such that its objective function is the aggregate welfare $mW^c[s_c, R(s_c, m), m]$. The fringe countries observe the leader's action s_c. Their 'aggregate' response is $s_f = R(s_c, m)$, and therefore the resultant welfare is $W^f[s_c, R(s_c, m), m]$ for each individual fringe country. Since the function W^c is inverse u-shaped and strictly concave in s_c (see Appendix B), there exists a unique solution to the coalition's optimization problem

$$\max_{s_c \in [0, mT^{-1}(0)]} mW^c[s_c, R(s_c, m), m]. \tag{17}$$

The Stackelberg equilibrium $[s_c^*, s_f^* = R(s_c^*, m)]$ is a point in the strategy space at which the best-reply function R of the fringe and an iso-welfare curve of the coalition are tangent.

In the sequel we will characterize the solution of (17), its relation to the BAU equilibrium and its dependence on the (exogenous) size of the coalition. We proceed in several steps beginning with the implications of an arbitrary action $s_c \in [0, mT^{-1}(0)]$ of the leader.

The coalition's anticipation of the fringe's reactions as driving force of outcomes The best-reply function of the fringe, (14), is of special interest, because all feasible outcomes necessarily satisfy that function. Accounting for R the coalition knows that its own emissions and those of the fringe are strategic substitutes. So it takes into consideration that if it reduces its emissions by the amount $\Delta s_c < 0$ [increases its emissions by the amount $\Delta s_c > 0$] total emissions will shrink [expand], but by less than $\left|\Delta s_c\right|$.[14] In the climate change literature this phenomenon is referred to as carbon leakage for the case $\Delta s_c < 0$. The leakage rate is usually expressed by $\left|R_{s_c}\right| \in]0,1[$. Since $R_{s_c m} > 0$ (Appendix A), the leakage rate is declining in the coalition size - which conforms to intuition and will turn out to drive the results. One can also show that[15]

$$e_c \gtreqless e_o \quad \Longleftrightarrow \quad \text{coalition} \left\{ \begin{array}{c} \text{imports} \\ \text{doesn't trade} \\ \text{exports} \end{array} \right\} \text{fuel.} \tag{18}$$

If e_c is kept constant, total emissions are rising in m, because the fringe's responding emission reduction falls short of the coalition's emission increase. Moreover, if m is kept constant, total emissions are rising in e_c because the leakage rate is positive but less than one.[16] Finally we note that the increase in total emissions resulting from a given increase in the coalition countries' emissions is the larger, the larger is the coalition size.[17] That is, large coalitions are more effective in curbing total emissions, because the leakage rate is declining in the coalition size.

Coalition size, equilibrium emissions and welfares, and their relation to BAU According to our previous analysis the entire Stackelberg equilibrium allocation is uniquely determined by - and varies with - the coalition size. To formalize that observation it is

[14] That follows from $s_c + s_f = s_c(1 + R_{s_c}) + R(0, m)$, with $(1 + R_{s_c}) \in]0, 1[$ because of $R_{s_c} \in]-1, 0[$.

[15] (18) follows from $x_i^s = T\left(\frac{s_c + s_f}{n}\right), R_{s_c} \in]-1, 0[$ and $\operatorname{sign}(e_c - e_o) = \operatorname{sign}(s_c + s_f - n e_o)$.

[16] The total differential of $s_c + s_f = s_c(1 + R_{s_c}) + R(0, m)$ reads $\mathrm{d}(s_c + s_f) = \left[\underbrace{(1 + R_{s_c})e_c + s_c R_{s_c m}}_{(+)}\right] \cdot \mathrm{d}m + \underbrace{m(1 + R_{s_c})}_{(+)} \cdot \mathrm{d}e_c$. Here we treat m as a real number in $[1, n]$ for analytical convenience although we will keep in mind that in real-world coalitions m is an integer in the set $\{1, \ldots, n\}$.

[17] Formally this follows from $\frac{\partial^2(s_c + s_f)}{\partial e_c \partial m} = (1 + R_{s_c}) + m R_{s_c m} = \frac{\{[1-(n-m-1)R_{s_c}]^2 - m R_{s_c}\}(1+R_{s_c})}{[1-(n-m-1)R_{s_c}]^2} > 0$.

analytically convenient to introduce the notation

$$
\begin{aligned}
&e_c^* = \mathcal{E}^c(m), \quad e_f^* = \mathcal{E}^f(m), \\
&\mathcal{W}^c(m) := W^c[m\mathcal{E}^c(m), (n-m)\mathcal{E}^f(m), m] \quad \text{and} \\
&\mathcal{W}^f(m) := W^f[m\mathcal{E}^c(m), (n-m)\mathcal{E}^f(m), m],
\end{aligned}
$$

and to consider the interval $[0, n]$ to be the domain of all these functions. A first but important result is the following proposition proved in Appendix C.

Proposition 1. *The Stackelberg equilibrium with the coalition of size $\tilde{m} \in [1, n]$ coincides with the non-cooperative BAU equilibrium, if and only if*

$$
\tilde{m} := \frac{(\alpha + b + n)n^2}{\alpha(2n-1) + n^2(b+1)} > 1. \tag{19}
$$

Proposition 1 specifies the link between Stackelberg equilibria and the BAU equilibrium. For the coalition it is optimal to choose the BAU emissions $e_c^* = e_o$ (leading to $e_f^* = e_o$), if and only if it has $\tilde{m}$ members. $\mathcal{E}^c(m) \neq e_o$ and $\mathcal{W}^c(m) \geq \mathcal{W}^c(\tilde{m})$ for all $m \neq \tilde{m}$ follows immediately from the observations that the benchmark coalition size $\tilde{m}$ is unique and that for any given m the coalition can always choose the emission cap $e_c = e_o$ which then leads to the BAU equilibrium. According to (19) $\tilde{m}$ varies with the model parameters and that feature will turn out to be of special interest below.

With the coalition size $\tilde{m}$ as a benchmark we are able to shed more light on the links between coalition size and deviations from BAU of emissions and welfare levels in Stackelberg equilibria. Suppose, the coalition of size m chooses the strategy $s_c = me_o$ and thus implements the BAU equilibrium. For all coalitions of size $m \neq \tilde{m}$ the strategy $s_c = me_o$ is clearly feasible but sub-optimal. Hence $MWC_o(m) \neq 0$ for all $m \neq \tilde{m}$, where $MWC_o(m)$ is a shorthand for the "**M**arginal (aggregate) **W**elfare of a **C**oalition of size $m \neq \tilde{m}$ evaluated at the 'BAU equilibrium strategy' $s_c = me_o$ ". We prove in the Appendix C that[18]

$$
MWC_o(m) \gtreqless 0 \quad \Longleftrightarrow \quad m \lesseqgtr \tilde{m}. \tag{20}
$$

For the interpretation of (20) we invoke our result from the Appendix C that the coalition's marginal *consumption* welfare in BAU is independent of the coalition size, so that variations in *total* marginal welfare result from variations in the coalition's marginal *climate* welfare exclusively. Hence, total BAU emissions ne_o are considered suboptimally large by large coalitions ($m > \tilde{m}$) and suboptimally small by small coalitions ($m < \tilde{m}$).[19] We combine

[18] Throughout the paper the subscript "o" refers to the BAU equilibrium.

[19] The reason for that differential effect is our finding that the effectiveness of curbing total emissions is increasing in the coalition size because the leakage rate declines with the coalition size. See footnote 17.

the information of (14) and (20) with the properties of $W^c[me_c, R(me_c, m), m]$ specified in Appendix B to conclude:

$$\mathcal{E}^c(m) \gtreqless e_o \quad \text{and} \quad \mathcal{E}^f(m) \lesseqgtr e_o \quad \Longleftrightarrow \quad m \lesseqgtr \tilde{m}. \tag{21}$$

The rationale of (21) is straightforward in view of footnote 17. In case of $m < \tilde{m}$ the leakage rate is high so that the coalition achieves a small reduction in total emissions only, if it reduces e_c. That makes total emission reductions very expensive. If, instead, the coalition relaxes rather than tightens the emission cap e_c, the resulting increase in total emissions is small owing to the high leakage rate, but the gain in consumption welfare is relatively large. Mirror symmetric arguments apply to the case $m > \tilde{m}$. Since leakage rates are always less than unity,

$$[m\mathcal{E}^c(m) + (n-m)\mathcal{E}^f(m)] \gtreqless ne_o \quad \Longleftrightarrow \quad m \lesseqgtr \tilde{m} \tag{22}$$

follows from (21). According to (22), small coalitions ($m < \tilde{m}$) do not mitigate but rather aggravate climate damage compared to BAU. It would therefore be more appropriate to call such coalitions 'anti-climate coalitions' rather than a 'climate coalitions'.

Turning to the coalition countries' welfare, note first that (20) implies that $\mathcal{W}^c(m)$ is strictly greater than $\mathcal{W}^c(\tilde{m})$ for all $m \neq \tilde{m}$. Hence the function W^c attains its absolute minimum[20] at $m = \tilde{m}$. Moreover, we verify in the Appendix D that

$$\left\{\begin{matrix} \mathcal{W}^c(m) > W_o > \mathcal{W}^f(m) \\ \mathcal{W}^c(m) = W_o = \mathcal{W}^f(m) \\ \mathcal{W}^f(m) > \mathcal{W}^c(m) > W_o \end{matrix}\right\} \quad \Longleftrightarrow \quad m \left\{\begin{matrix} < \\ = \\ > \end{matrix}\right\} \tilde{m} \tag{23}$$

with $W_o := \mathcal{W}^c(\tilde{m}) = \mathcal{W}^f(\tilde{m})$. In case of $m < \tilde{m}$ the coalition finds it beneficial to expand own emissions above BAU level which induces the fringe countries to tighten their emission caps. The opportunity costs of that policy on the part of fringe countries is consumption foregone. The consumption welfare loss combined with the reduction in climate welfare pushes the fringe countries' total welfare below BAU level. Thus, the coalition free rides on the fringe countries' mitigation efforts. In case of $m > \tilde{m}$ the roles of both groups are reversed. Now the fringe countries free ride on the coalition's mitigation policy, which is their expected role, and the fringe countries benefit on two margins: Their consumption welfare rises compared to BAU as well as their climate welfare. A general principle appears to be that countries with laxer emission regulation have higher welfare levels. So far, we summarize our results in

Proposition 2. *Consider the transition from BAU to the Stackelberg equilibrium. The shift of*

[20] In our numerical calculations $\mathcal{W}^c$ will turn out to be u-shaped as shown in Figure 2 below.

(i) the coalition country's emissions is characterized in (21);

(ii) total emissions is characterized in (22);

(iii) the coalition country's and fringe country's welfare is characterized in (23).

The results of Proposition 2 provide interesting information about the relations between the coalition size, the BAU equilibrium and the Stackelberg equilibrium. However, as the functions $\mathcal{E}^j$ and $\mathcal{W}^j$ for $j = c, f$ depend on m in a very complex way, their curvature cannot be specified analytically.

Numerical example To make progress we resort to a numerical example, referred to as Example 1,[21] with the parameter values $a = 100$, $b = 20$, $\bar{x} = 12$, $\alpha = 1000$ and $n = 10$. The Figures 1 and 2 display the pertaining graphs of the functions $\mathcal{E}^h$ and $\mathcal{W}^h$ for $h = c, f$ and the curves of aggregate emissions and welfares, respectively. Observe that (21), (22) and (23) are satisfied in these figures. The new information conveyed by Example 1 is that the function $\mathcal{E}^c$ [$\mathcal{E}^f$] is strictly decreasing [increasing] and that total emissions $m\mathcal{E}^c + (n-m)\mathcal{E}^f$ are strictly decreasing in m.[22]

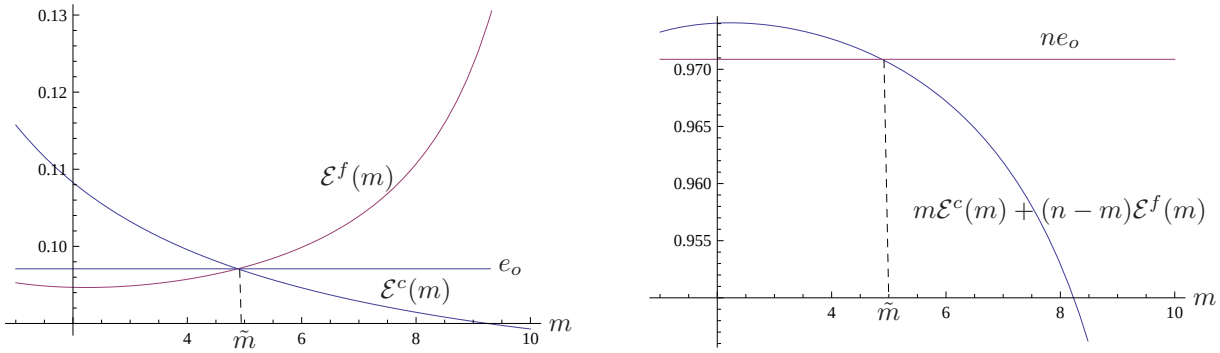

Figure 1: Emissions caps and total emissions in Example 1

According to the left panel of Figure 2, the (total) welfare of coalition countries is u-shaped with its unique minimum at $m = \tilde{m}$, whereas $\mathcal{W}^f$ is strictly increasing in m. The surprising feature of the right panel of Figure 2 is not that the world welfare rises in m but that for all $m < \tilde{m}$ the world welfare falls short of its BAU level. The coalition of size $m < \tilde{m}$

[21] We cannot generalize our findings from Example 1 by induction, of course. Yet we have run several other examples, e.g. the Example 2 specified by the parameter values $a = 1000$, $b = 2000$, $\bar{x} = 12$, $\alpha = 500000$, and $n = 100$ (to be considered in the next section). The graphs corresponding to all examples under scrutiny turned out to be qualitatively the same as those in the Figures 1, 2 and 3 which is why we restrict the graphical presentation to Example 1.

[22] We consider as negligible that the functions $\mathcal{E}^f, \mathcal{W}^f, m\mathcal{W}^c + (n-m)\mathcal{W}^f$ and $m\mathcal{E}^c + (n-m)\mathcal{E}^f$ are slightly non-monotone for $m < 2$.

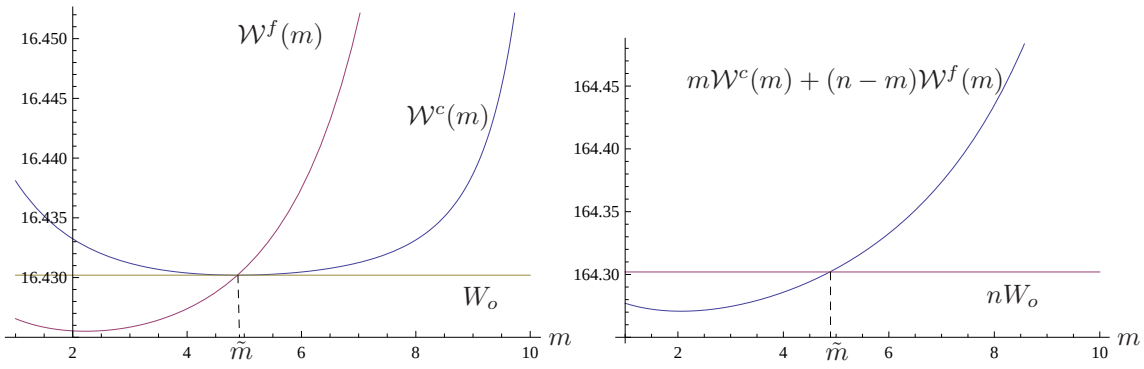

Figure 2: Welfare and aggregate welfare in Example 1

clearly succeeds in raising its welfare above BAU level by increasing the climate damage at the expense of the fringe countries. As the latter engage in costly mitigation to keep the increase in total emissions small, they suffer a welfare loss compared to BAU (left panel of Figure 2) which is even larger than the coalition's welfare gain.

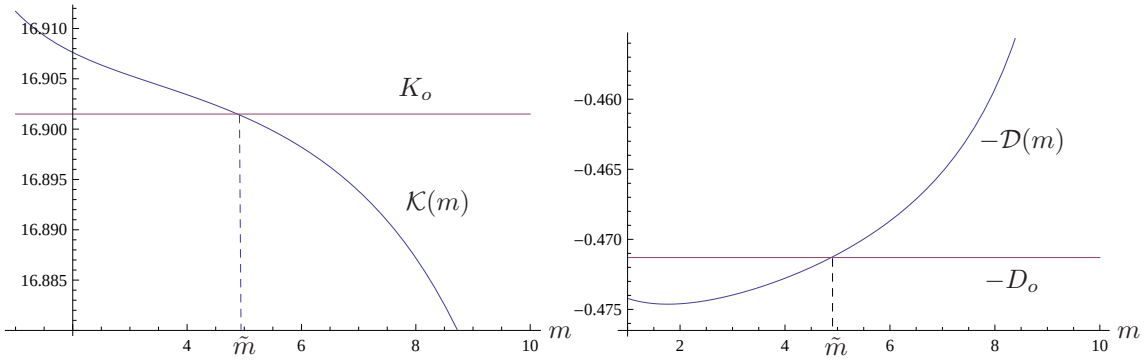

Figure 3: Consumption welfare $[\mathcal{K}(m)]$ and climate welfare $[-\mathcal{D}(m)]$ of the coalition in Example 1

Figure 3 decomposes the (total) welfare of a coalition country into its consumption welfare (curve $\mathcal{K}(m)$) and climate welfare (curve $-\mathcal{D}(m)$). Figure 3 illustrates that owing to the high leakage rate in case of $m < \tilde{m}$, the coalition finds it advantageous to sacrifice, compared to BAU, some climate welfare for additional consumption welfare. Conversely, if $m > \tilde{m}$, the coalition is more effective in reducing total emissions and therefore benefits from shifting away from consumption welfare toward higher climate welfare.

3.2 Self-enforcing IEAs

In the preceding Section 3.1 we have presupposed the presence of a climate coalition, and our focus has been on characterizing the Stackelberg equilibrium and its dependence on the exogenous coalition size m. Now we turn to the issue of coalition stability. Since

supranational authorities for the effective enforcement of agreements are not available, IEAs will not prevail unless they are self-enforcing in the sense that no signatory has an incentive to defect (*internal stability*) and no non-signatory has an incentive to sign the agreement (*external stability*).[23] In formal language, an IEA with $m \in \{1, \ldots, n\}$ signatories is said to be self-enforcing or stable if it satisfies the internal stability condition

$$\mathcal{W}^c(m) \geq \mathcal{W}^f(m-1) \tag{24}$$

and the external stability condition

$$\mathcal{W}^f(m) \geq \mathcal{W}^c(m+1). \tag{25}$$

With the distinction between the membership $m \in \{1, \ldots, n\}$ of real-world IEAs and the real-number approximation $m \in [1, n]$ in mind we find that if a self-enforcing IEA with $m^* \in \{1, \ldots, n\}$ signatories exists, then $m^* > \tilde{m}$. To verify that claim, note that $\mathcal{W}^c(m) > W_o > \mathcal{W}^f(m)$ for all $m < \tilde{m}$ from (23) implies $\mathcal{W}^f(m) < \mathcal{W}^c(m+1)$. So the external stability condition is violated for all $m \in \{1, \ldots, n\}$ with $m < \tilde{m}$. If $\tilde{m}$ happens to be an integer, the coalition of size $\tilde{m}$ is not stable either, because fringe countries have still an incentive to join the coalition ($W^f(\tilde{m}) < W^c(\tilde{m}+1)$). Hence all those coalitions fail to be stable that push up total emissions above BAU level. The downside of our finding "$m^* > \tilde{m}$, if m^* exists" is that it leaves open whether m^* exists, and if so, how large the positive difference $(m^* - \tilde{m})$ is. Unfortunately, we have not been able to answer the existence question analytically. We therefore resort to examining the stability conditions (24) and (25) for the numerical Examples 1 and 2 introduced in the previous Section 3.1.

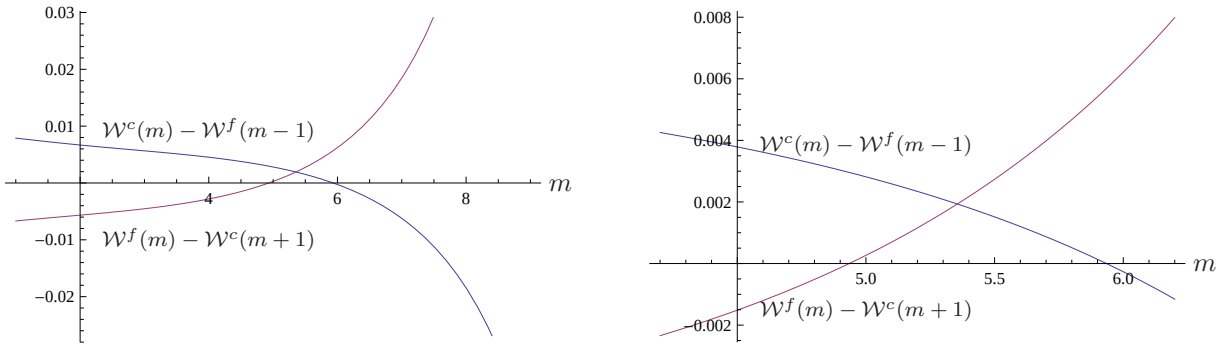

Figure 4: Stability in Example 1 ($\tilde{m} = 4.881, m^* = 5$)

[23] This notion of self-enforcement or stability was originally introduced by D'Asprement et al. (1983) in the context of cartel formation.

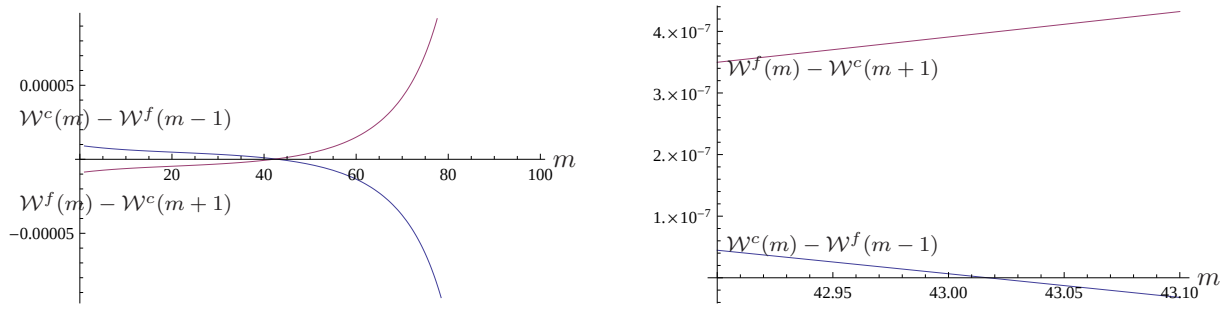

Figure 5: Stability in Example 2 ($\tilde{m} = 42.013, m^* = 43$)

The Figures 4 and 5 present the graphs of the functions $\mathcal{W}^c(m) - \mathcal{W}^f(m-1)$ and $\mathcal{W}^f(m) - \mathcal{W}^c(m+1)$ for the Examples 1 and 2, and their right panels exhibit an enlarged detail of the relevant domain. In both cases there is one and only one interval of coalition sizes in which both functions take on non-negative values (thus satisfying (24) as well as (25)), and this interval contains one and only one integer, $m^* = 5$ in Example 1 and $m^* = 43$ in Example 2. Moreover, in both cases the stable coalition size m^* is the smallest integer greater than $\tilde{m}$ such that between 40% and 50% of all countries are members of the stable coalition. That contrasts sharply with the result of Rubio and Ulph (2006) and Diamantoudi and Sartzetakis (2006) according to whom the number of signatories in self-enforcing IEAs (in a world without trade) is small in the parameter sub-space securing positive equilibrium emissions.[24]

We carried out a number of examples in addition to the Examples 1 and 2 and their modifications in the Tables 1 and 2 below and reached the unequivocal result that for every parameter constellation securing positive equilibrium emissions there exists a unique self-enforcing IEA whose coalition size m^* is the smallest or second smallest integer larger than $\tilde{m}$ from (19). Thus it is clear from our comments on Proposition 1 that the allocation of Stackelberg equilibria with self-enforcing IEAs is approximately the same as in BAU; the climate damage is only slightly lower and the coalition countries' welfare is only slightly higher than in BAU, while the welfare gain of fringe countries is greater than that of coalition countries.

As we found that m^* is very close to $\tilde{m}$ in all of our examples we assess the determinants of the size of m^* by investigating the determinants of $\tilde{m}$. Recall that according to (19), $\tilde{m}$ depends on the size of the parameters α, b and n. To examine how $\tilde{m}$ varies with α, we differentiate (19) with respect to α and obtain

$$\frac{\mathrm{d}\tilde{m}}{\mathrm{d}\alpha} = \frac{n^2(n-1)[b(n-1)-n]}{[\alpha(2n-1)+n^2(b+1)]^2} \gtreqless 0 \quad \Longleftrightarrow \quad b \gtreqless \frac{n}{n-1}. \tag{26}$$

[24]It is straightforward from the left panels of the Figures 4 and 5 that the equilibrium emissions $\mathcal{E}^f(m^*)$ and $\mathcal{E}^c(m^*)$ are strictly positive.

For α converging to infinity we find $\lim_{\alpha\to\infty}\tilde{m} = \frac{n^2}{2n-1}$.

α	1	10	50	100	500	1000	1450	∞
$\tilde{m}$	1.46	1.75	2.62	3.25	4.57	4.88	4.99	5.26
m^*	3	3	3	4	5	5	6	6

Table 1: Variations of α in Example 1 ($n = 10$)

α	10^3	10^4	10^5	$5 \cdot 10^5$	10^6	10^7	∞
$\tilde{m}$	1.53	5.50	25.58	42.01	45.75	49.76	50.25
m^*	3	6	26	43	46	50	51

Table 2: Variations of α in Example 2 ($n = 100$)

According to (26) the comparative static effect of α depends on the size of b. The values of b and n chosen in the Examples 1 and 2 satisfy $b > n/(n-1)$ such that $\tilde{m}$ is increasing in α and converges toward $n^2/(2n-1)$ from below. This is confirmed by the numerical examples listed in the Tables 1 and 2. If $b < n/(n-1)$, $\tilde{m}$ is decreasing in α and converges toward $n^2/(2n-1)$ from above. That is, for $b < n/(2n-1)$ and α sufficiently small, equation (19) allows for high levels of $\tilde{m}$, even for $\tilde{m} = n$ (grand coalition). However, the non-negativity constraint for emissions turns out to be violated for low values of α (and $b < n/(n-1)$). We did not find any numerical example of Stackelberg equilibria exhibiting both non-negative emissions and stable coalition sizes larger than $\frac{n^2}{2n-1}$. Hence under the condition of positive equilibrium emissions the maximum share of countries in a stable coalition, $100\,m^*/n$, appears to be slightly higher than 50%. We need to emphasize, however, that there are various examples in the Tables 1 and 2 in which the share $100\,m^*/n$ is much smaller than 50%. It is also worth noting that in all cases but one calculated in the Tables 1 and 2 m^* is the smallest or second smallest integer larger than $\tilde{m}$.

The role the parameter α plays in the formation of self-enforcing IEAs calls for an economic interpretation. To keep focussed we restrict our attention to the set of parameters satisfying $b > n/(n-1)$ and define the fuel extraction costs, expressed in units of good X, as

$$C(e_i^s;\alpha) := T(0) - T(e_i^s) = \frac{\alpha}{2}(e_i^s)^2 \tag{27}$$

Those extraction costs are obviously progressively increasing such that increasing α corresponds to increasing marginal extraction costs which increases the size of stable coalitions in turn. The lower and the less progressive the extraction costs, the smaller is the size of the stable coalition. We summarize our results in

Proposition 3. *Under the conditon of positive equilibrium emissions there exist self-enforcing IEAs that are characterized as follows:*

(i) If $b > n/(n-1)$, then the stable coalition size m^* increases in the parameter α such that as many as (slightly more than) 50% of all countries may be members of a stable coalition.

(ii) The number of countries in the self-enforcing IEAs is the smallest or the second smallest integer m^* larger than $\tilde{m}$ from (19). Therefore the corresponding Stackelberg equilibrium allocation differs only slightly from the allocation in the scenario of global non-cooperation (BAU).

We are aware of the limited scope of Proposition 3 because it is based on numerical examples. Nonetheless, the unequivocal result of the calculations we conducted suggests that the messages of Proposition 3 are more general. Proposition 3(i) gives support to the expectation that international trade may lead to rather large stable coalitions. That appears to be good news for supporters of strong climate damage mitigation action, if large stable coalitions promise to bring about reductions of global emissions that are larger by an order of magnitude than in BAU and hopefully not too far away from the socially optimal allocation. Unfortunately, Proposition 3(ii) shatters that expectation. Our numerical calculations rather suggest that stable coalitions reduce world emissions only insignificantly compared to BAU emissions. To the extent that this result is general - which we are not able to prove analytically - the highly inconvenient implication is that efforts to reach a self-enforcing IEA do not pay.

Proposition 3(ii) calls for explanation and economic intuition. It is clear from the conditions (24) and (25) that the stability of coalitions depends on the properties of the functions $\mathcal{W}^c$ and $\mathcal{W}^f$. In the left panel of Figure 2 we see that $\mathcal{W}^f(m) - \mathcal{W}^c(m)$, the *vertical* difference between the welfare curves $\mathcal{W}^f$ and $\mathcal{W}^c$, is zero for $m = \tilde{m}$ and positive for all $m > \tilde{m}$. That difference can be interpreted as the free-rider advantage of fringe countries over coalition countries. In our Example 1 that free-rider advantage grows with the coalition size suggesting that the incentive to leave the coalition increases and the incentive to join declines with the coalition size. However, the defining criterion for coalition stability is the *horizontal* rather than the vertical distance between the welfare curves $\mathcal{W}^f$ and $\mathcal{W}^c$. To be more specific, we introduce the function

$$H : [\tilde{m}, n] \longrightarrow \mathbb{R}_+, \text{ where } \quad h = H(m), \quad \text{if and only if } \quad \mathcal{W}^f(m-h) = \mathcal{W}^c(m).$$

$H(m)$ measures the *horizontal* distance between the welfare curves $\mathcal{W}^f$ and $\mathcal{W}^c$ at the level $\mathcal{W}^c(m)$ above the horizontal m-axis. Unfortunately, analytical complexity prevents us from

determining the curvature of H in our parametric model. Figure 6 shows for Example 1 that the function H is strictly increasing in m on the relevant part of its domain.[25] As we found that kind of curvature of H in all of our examples, it appears to be robust.

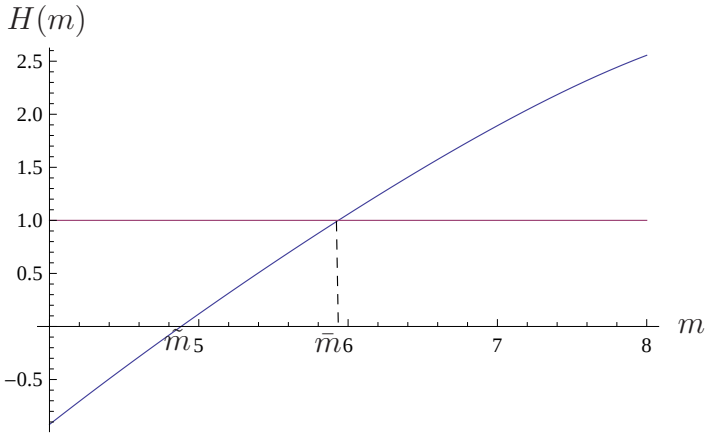

Figure 6: Function H in case of Example 1

From Figure 6 and the definition of coalition stability follows that there is a unique value $\bar{m} > \tilde{m} + 1$ satisfying $H(\bar{m}) = 1$ and a stable coalition[26] of size $m^* \geq \tilde{m}$ satisfying $m^* \in [\bar{m} - 1, \bar{m}]$. Obviously, the (positive) difference $m^* - \tilde{m}$ is the smaller, the smaller is $\bar{m}$, and $\bar{m}$ is the smaller, the larger is the slope of function H. In other words, the faster $H(m)$ grows in m, the smaller is the coalition size at which coalitions become internally instable and the closer to $\tilde{m}$ moves the size m^* of the stable coalition. Hence the slope H_m measures how fast the coalition countries' incentives to defect increase with increasing coalition size. Since $\bar{m} > \tilde{m} + 1$,[27] $\bar{m} \in]\tilde{m} + 1, \tilde{m} + 2[$ is as close as $\bar{m}$ can move toward $\tilde{m}$, and, in fact, $\bar{m} \in]\tilde{m} + 1, \tilde{m} + 2[$ holds in all of our numerical examples. As an implication, coalition countries defect soon after the coalition size exceeds the BAU coalition size $\tilde{m}$. It is straightforward to show that the stable coalition size m^* is the smallest or second smallest integer larger than $\tilde{m}$, if and only if $\bar{m} \in]\tilde{m} + 1, \tilde{m} + 2[$.

Our preceding arguments aimed at identifying the driving forces of coalition stability. For further interpretation we define the functions $\Omega^j : [\tilde{m}, n] \longrightarrow \mathbb{R}_+$, $j = c, f$ by

$$\Omega^c(m) = \omega_o + \frac{\omega_1}{2}(m - \tilde{m})^2 \quad \text{and} \quad \Omega^f(m) = \omega_o + \omega_2(m - \tilde{m}) + \frac{(\omega_1 + \omega_3)}{2}(m - \tilde{m})^2, \quad (28)$$

where the parameters $\omega_o, \omega_1, \omega_2$ and ω_3 are assumed to satisfy $\omega_o = \mathcal{W}^f(\tilde{m}) = \mathcal{W}^c(\tilde{m})$, $\omega_1 > 0$, $\omega_2 = \mathcal{W}^f_m(\tilde{m}) > 0$ and $\omega_3 \geq 0$. By construction, the functions Ω^c and Ω^f approximate the functions $\mathcal{W}^c$ and $\mathcal{W}^f$ (see Appendix E). Taking advantage of that approximation we prove in the Appendix E

[25] The relevant part of the domain is $[\tilde{m}, \breve{m}]$, where $\breve{m} = \bar{m} + H(\breve{m})$ and where $\bar{m}$ is defined by $H(\bar{m}) = 1$.
[26] The coalition of size $m^* > \tilde{m}$ is stable, if and only if $H(m^*) \leq 1$ and $H(\breve{m}) = \breve{m} - m^* \geq 1$.
[27] $\bar{m} > \tilde{m} + 1$ because $H(m) \in [0, 1[$ for all $m \in [\tilde{m}, \tilde{m} + 1[$.

Proposition 4. *Approximate the functions $\mathcal{W}^c$ and $\mathcal{W}^f$ on the sub-domain $[\tilde{m}, n]$ by the functions Ω^c and Ω^f defined in (28).*

(i) There exists a stable coalition of size $m^* > \tilde{m}$ and m^* is unique, in general.[28]

(ii) Ceteris paribus, the difference $m^* - \tilde{m}$ is the smaller,

- *the slower the coalition countries' welfare increases with the coalition size ($\omega_1 \downarrow$) due to tight domestic emission caps and leakage-retarded climate damage reduction;*
- *the faster the fringe countries' welfare increases with the coalition size ($\omega_2 \uparrow$ and $\omega_3 \uparrow$) due to lax domestic emission caps (reflecting carbon leakage) and free rides on climate damage reduction.*

(iii) The size of the stable coalition is either the smallest or the second smallest integer weakly larger than $\tilde{m}$, if and only if $3\omega_1 < 2\omega_2 + \omega_3$.

Proposition 4(i) confirms for the 'auxiliary' functions Ω^c and Ω^f our findings reported in the context of Figure 6. Proposition 4(ii) identifies the parameters ω_1, ω_2 and ω_3 as determinants of the size of the difference $m^* - \tilde{m}$ and links these parameters to their economic impacts in a straightforward way. Proposition 4(iii) clarifies that the stable coalition size m^* will be the smallest or second smallest integer larger than $\tilde{m}$, if and only if the *relative* difference between the welfare increases in the coalition size of fringe and coalition countries is sufficiently large. Here the change in the vertical distance between the functions Ω^f and Ω^c determines the size of the difference $m^* - \tilde{m}$ because in case of positive parameters ω_1, ω_2 and ω_3 the vertical as well as the horizontal difference is increasing in m.

4 On the role of international trade

Up to now we have dealt with a world economy characterized by the four parameters $(a, \alpha, b, n) \in \mathbb{R}^4_{++}$ in the regime of free trade. The straightforward way of improving our understanding of the role of international trade for the formation of self-enforcing IEAs is to compare the results derived in the free-trade model with those of the autarky scenario in the otherwise unchanged model. The only substantive modification of the model (1) - (9) is to replace (4) by

$$x_i^s = x_i^d \quad \text{and} \quad e_i^s = e_i^d \qquad i = 1, \ldots, n, \tag{29}$$

which simply turns the world markets for good X and fossil fuel into domestic markets. Good X can still be taken as numéraire ($p_{xi} = 1$ for $i = 1, \ldots, n$) but (5) is now replaced

[28] We show in the proof (Appendix E) that in exceptional cases there are two stable coalitions.

by $p_{ei} = -T'(e_i)$ for $i = 1, \ldots, n$. With these changes the welfare of country i is

$$W^i(e_1, \ldots, e_n) = V(e_i) + T(e_i) - D\left(\sum_j e_j\right) \tag{30}$$

for the general functions (1) and (2) and

$$W^i(e_1, \ldots, e_n) = ae_i - \frac{\check{b}}{2}e_i^2 + \bar{x} - \frac{1}{2}\left(\sum_j e_j\right)^2 \tag{31}$$

with $\check{b} := b + \alpha$ for the parametric functions (3).

The comparison of (12) and (30) subject to (3) shows that the switch from free trade to autarky turns the economy (a, α, b, n) into the economy $(a, \check{\alpha} = 0, \check{b} = b + \alpha, n)$. The latter obviously has the structure of the basic model of the coalition formation literature in which production and international trade is not modeled.[29] Thus our free-trade versus autarky comparison is also a comparison between the basic model and our trade model. In the following we carry out that comparison in several steps.

To begin with, the BAU equilibria of the economy (a, α, b, n) with and without trade coincide, because comparative advantage is absent if identical countries are treated equally. Moreover, along the lines of the proof of $\tilde{m}$ in (19) one can show that the coalition size[30]

$$\tilde{m}_a := \frac{\check{b} + n}{\check{b} + 1} \tag{32}$$

for which the Stackelberg equilibrium (in case of real-number coalitions) is equal to the BAU equilibrium in the economy (a, α, b, n). The comparison of (32) with (26) readily yields $\tilde{m}_a < \tilde{m}$.

Since in the regime of autarky the model of the present paper coincides with the basic model of the coalition formation literature, we can invoke the results of Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006). They show that "... restricting parameter values to guarantees interior solutions is a sufficient condition to get stable IEAs with a small number of signatories ..." (Rubio and Ulph, 2006, p. 236). Diamantoudi and Sartzetakis focus exclusively, as we do, on subsets of parameters leading to positive equilibrium emissions and find that stable IEAs have at most four signatories even if the total number of countries is large. Rubio and Ulph (2006) consider a larger parameter space and introduce non-negativity constraints on emissions. For a subset of parameter values which guarantees

[29] See e.g. Finus (2001, equation (3.1)). Diamantoudi and Sartzetakis (2006, equation (1)) as well as Rubio and Ulph (2006, equation (1)) restrict their analysis to the parametric version (31) of the *basic model*.

[30] In the sequel the autarky regime is indicated by the super- or subscript a.

interior solutions they find that the maximum stable coalition size is three.[31]

To sum up, as long as solutions with non-positive emissions are ruled out, we get stable IEAs with a small number of signatories in the autarky scenario (= basic model) irrespective of the total number of countries. That result clearly is in stark contrast to our finding in the free-trade model of Section 3 where we have identified stable coalitions much larger than in the autarky model.

Regarding the comparison of free trade and autarky, we also want to know how effective the stable coalition is in reducing world emissions below BAU emissions. Rubio and Ulph (2006) focus on a parameter space that secures positive equilibrium emissions (ibidem, footnote 16) and point out that $m_a^* \leq 3$ (ibidem Corollary 2). However, they do not address the size of the difference $m_a^* - \tilde{m}_a$. Diamontoudi and Sartzetakis (2006) find that the welfare of the signatories is very close to its lowest value when the IEA is stable but they do not link that observation to the BAU scenario. Analogous to the result $m^* > \tilde{m}$ in Section 3.2, it is straightforward to establish that in autarky the size of a self-enforcing IEA satisfies $m_a^* > \tilde{m}_a$. Moreover, the Appendix G proves that $m_a^* - \tilde{m}_a \leq 2$ for all economies in the parameter space considered in Rubio and Ulph (2006). We summarize these findings in

Proposition 5. *Consider the world economy without international trade for a parameter space introduced by Rubio and Ulph (2006) that secures positive equilibrium emissions.*

(i) Then our model coincides with the 'basic model' studied e.g. by Diamantoudi and Sartzetakis (2006) and Rubio and Ulph (2006).

(ii) Then the size m_a^* of self-enforcing IEAs is the smallest or second smallest integer larger than $\tilde{m}_a$ from (32), and at most equal to 3.

(iii) The emission caps of the signatories of the self-enforcing IEA are only slightly tighter than the emission cap in the BAU equilibrium.

The remainder of Section 4 serves to explain the differences in outcome between the scenarios of autarky and free trade. Since in both cases the stable Stackelberg equilibrium is close to BAU, the reasons for "$m_a^* - \tilde{m}_a > 0$ but small" are the same in qualitative terms as those for "$m^* - \tilde{m} > 0$ but small" discussed in Section 3.2. Hence we can restrict our focus on explanations for $\tilde{m} > \tilde{m}_a$.

[31] Barrett (1994) shows that there are parameter constellations for which the self-enforcing IEA may attain any size from very small to the grand coalition. That finding is not at variance with our results because Diamantoudi and Sartzetakis (2006) convert Barrett's approach into the *basic model* of type (31) and show that in Barrett's framework self-enforcing IEAs consist of no more than four countries on the set of parameters leading to positive equilibrium emissions.

One way to highlight the reason for $\tilde{m} > \tilde{m}_a$ is to invoke the fuel extraction costs $C(e_i^s;\alpha)$ from (27). For $\alpha \to 0$ extraction costs $C(e_s^i;\alpha)$ tend to zero and fuel becomes a free good. In that case there is no need and no role for international trade anymore because the outcomes are the same under open and closed borders. Thus we can interpret the economy (a,α,b,n) in the regime of autarky - as well as the basic model of the literature - as the 'polar case' of a free-trade economy with zero fuel extraction costs. In that perspective the absence of extraction costs is the reason for $\tilde{m}_a < \tilde{m}$.

In the autarkic economy (a,α,b,n) the fringe countries' best-reply function is characterized by the first-order condition $V'(e_f)+T'(e_f)-D'[me_c+(n-m)e_f]=0$ which implicitly determines the aggregate best-reply function of the fringe, denoted $s_f = R^a(s_c,m)$. It is straightforward to show that the function R^a exhibits the same qualitative properties as the function R from (14) such that (20) as well as the results in the Appendixes A and B carry over to the autarky regime. Likewise, Proposition 1 still holds, when we replace $\tilde{m}$ by $\tilde{m}_a$, and it is true that, if it exists, the size of the stable coalition in autarky, m_a^*, is larger than $\tilde{m}_a$. We prove in the Appendix F that $\left|R_{s_c}\right| > \left|R_{s_c}^a\right|$. That important quantitative difference between both regimes means that the leakage rate is larger in the free-trade regime than in autarky.[32] As an immediate consequence of (31) the marginal (aggregate) welfare of coalition countries evaluated at BAU (defined in Appendix C) is lower under free trade than under autarky, formally $MWC_o(m) < MWC_o^a(m)$. Since $\tilde{m}$ and $\tilde{m}_a$ are determined by $MWC_o(\tilde{m})=0$ and $MWC_o^a(\tilde{m}_a)=0$, respectively, we infer from (20) and its analogue for autarky that $\tilde{m} > \tilde{m}_a$. Thus we have identified $\left|R_{s_c}\right| > \left|R_{s_c}^a\right|$ as a driver for $\tilde{m} > \tilde{m}_a$.

To further characterize the differences between autarky and free trade we consider Example 1 for autarky. The 'autarky functions' $\mathcal{E}^{ca}, \mathcal{E}^{fa}, \mathcal{W}^{ca}$ and $\mathcal{W}^{fa}$ turn out to exhibit the same qualitative properties (sign of slope, curvature) as the functions $\mathcal{E}^c, \mathcal{E}^f, \mathcal{W}^c$ and $\mathcal{W}^f$ in Example 1 with free trade. More specifically, the equivalences (21) - (23) (and hence Proposition 2) carry over to the autarky scenario, when the superscript a is attached to $\mathcal{E}^c, \mathcal{E}^f, \mathcal{W}^c$ and $\mathcal{W}^f$ and $\tilde{m}$ is replaced by $\tilde{m}_a$. In Example 1 the benchmark coalition size in autarky, $\tilde{m}_a = 1.009$, is much smaller than its free-trade counterpart $\tilde{m} = 4.881$. In the Figures 7 and 8 we illustrate the differences in outcome for the coalition countries under free trade and autarky and the dependence of these differences on the coalition size.

Recall that welfare consists of consumption welfare and climate welfare and that large coalitions are more effective in reducing total emissions. The right panel of Figure 7 il-

[32] Emissions of fringe and coalition are strategic substitutes under both free trade and autarky. But they are stronger strategic substitutes with trade than without. Copeland and Taylor (2005) reach the opposite conclusion in a model that differs substantially from ours - and even find conditions under which emissions of different countries turn into strategic complements when the borders are opened.

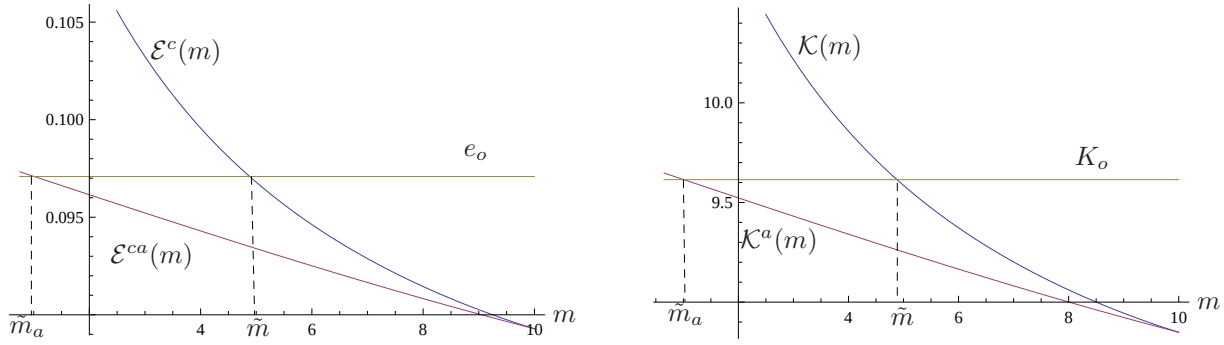

Figure 7: Autarky vs. free trade. Emissions and consumption welfare of coalition countries in Example 1.

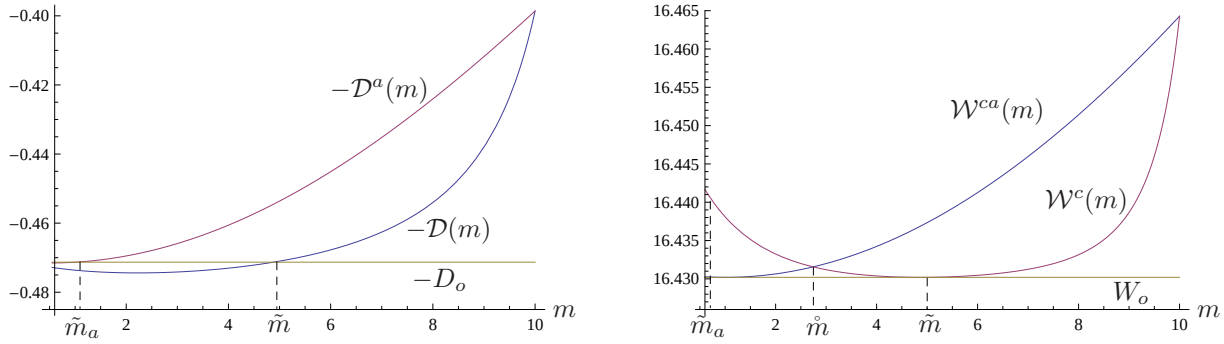

Figure 8: Autarky vs. free trade. Climate damage and total welfare of coalition countries in Example 1.

lustrates that with international trade the coalition countries enjoy much higher levels of consumption welfare than in autarky. For coalition sizes $m < \tilde{m}$ the coalition finds it beneficial to increase climate damage compared to BAU in order to raise consumption welfare through importing fossil fuel and exporting consumer goods. With closed borders these increases in consumption welfare are not attainable which makes climate damage reduction more attractive in autarky than under free trade. In autarky all countries necessarily consume what they produce whereas under free trade the coalition countries take advantage of decoupling consumption from production and thus raise their consumption welfare. Recall also that leakage rates under autarky are smaller than under free trade. Both effects imply that coalitions of given size set tighter emission caps (implying higher climate welfare as illustrated in the left panel of Figure 8) under autarky than under free trade, as can be seen in the left panel of Figure 7. Consequently, the size of the coalition choosing the BAU emission cap is larger under free trade than under autarky ($\tilde{m}_a < \tilde{m}$).

The left panel of Figure 7 shows that in both regimes the emissions of coalition countries are falling, that they are lower in autarky than in the trade regime, and that this difference tends to zero with m approaching n. The positive difference $\mathcal{E}^c(m) - \mathcal{E}^{ca}(m)$ is clearly

due to $\left|R_{s_c}\right| > \left|R^a_{s_c}\right|$ and the absence or presence of international trade. Taking the BAU consumption $[T(e_o), e_o]$ and the corresponding consumption welfare, K_o, as a benchmark, the right panel of Figure 7 illustrates the relationship between the levels of consumption welfare of coalition countries in BAU, K_o, in autarky, $K^a(m)$, and under free trade, $K(m)$. $K(m)$ is larger than $\mathcal{K}^a(m)$, because the fuel consumption $\mathcal{E}^c(m)$ is larger than $\mathcal{E}^{ca}(m)$ and because the coalition countries benefit from decoupling consumption and production. Due to the more stringent emission reduction in autarky, the climate welfare of all countries is higher in autarky than in free trade - as illustrated in the left panel of Figure 8.

As the coalition countries' consumption welfare is lower and their climate welfare is higher in autarky than in free trade, their net welfare change is ambiguous. More specific information on the comparison of total welfare $\mathcal{W}^{ca}(m)$ and $\mathcal{W}^c(m)$ provides the right panel of Figure 8. It shows that when moving from autarky to free trade, the coalition countries' climate welfare gain is overcompensated by their consumption welfare loss and the opposite holds for relatively large coalition sizes $m > \mathring{m}$.

5 Concluding remarks

The present paper reexamines the issue of self-enforcing international environmental agreements (IEAs) extending the *basic model* of the IEA literature introduced by Barrett (1994) and others to a general equilibrium framework with production, consumption and international trade. In model specifications yielding positive equilibrium emissions and with an IEA acting as Stackelberg leader we show

(a) that in stark contrast to the outcome of the *basic model* large stable IEAs may form,

(b) and that in all Stackelberg equilibria with a stable IEA the 'gains of cooperation' are negligible: Compared to the case of global non-cooperation the coalition countries' welfare gains as well as the climate damage reduction are very small.

While result (a) raises hopes for successful and effective cooperation in fighting climate change, result (b) thwarts these hopes because efforts of achieving effective mitigation through self-enforcing IEAs are futile irrespective of how large these IEAs are.

An interesting side result is that in the absence of international trade our model coincides with the *basic model* of the extant IEA literature, which means, in turn, that the basic model can be interpreted as a model of autarkic countries. We know from the literature that in the *basic model* the number of signatories of self-enforcing IEAs is very small, and we add the result that the allocation in the corresponding Stackelberg equilibrium does not

differ much from the business-as-usual allocation. As in the case of international trade, the coalition countries' welfare rises and the climate damage declines by a very small amount only.

Although our model has more 'economic' structure than the basic model we have kept it simple enough for the benefit of comparing it with the *basic model* and for the benefit of deriving informative results. As pointed out in the introduction the assumption of emissions being non-essential is not fully satisfactory for carbon emissions in the context of climate change mitigation. It is necessary and desirable to examine the outcome for the case of essential emissions, even if analytical results cannot be obtained anymore. More generally, one would want to check the robustness of results when economies are modeled in a more complex way, e.g. when fossil fuel is not only a final consumption good but also an intermediary industrial input. It is needless to say that while the assumption of symmetric countries is crucial for deriving meaningful (analytical) results, it abstracts from many real-world complexities which are severe barriers to reaching self-enforcing IEAs, and it therefore likely underestimates the difficulties of forming such agreements.

References

Barrett, S. (1994): Self-enforcing international environmental agreements. *Oxford Economic Papers* 46, 878-894.

Barrett, S. (1997): The strategy of trade sanctions in international environmental agreements. *Resource and Energy Economics* 19, 345-361.

Barrett, S. (1999): A theory of full international cooperation. *Journal of Theoretical Politics* 11, 519-541.

Barrett, S. (2001): International cooperation for sale. *European Economic Review* 45, 1835-1850.

Buchner, B., Carraro, C. and I. Cersosimo (2002): Economic consequences of the US withdrawal from the Kyoto/Bonn Protocol, *Climate Policy* 2, 273-292.

Carraro, C. and D. Siniscalco (1991): Strategies for the international protection of the environment. *CEPR discussion paper* 568.

Carraro, C. and D. Siniscalco (1993): Strategies for the international protection of the environment. *Journal of Public Economics* 52, 309-328.

Carbone, J.C., Helm, C. and T.F. Rutherford (2009): The case for international emission

trade in the absence of cooperative climate policy. *Journal of Environmental Economics and Management* 58, 233-263.

Copeland, B.R. and M.S. Taylor (2005): Free trade and global warming: a trade theory view of the Kyoto protocol. *Journal of Environmental Economics and Management* 49, 205-234.

D'Aspremont, C., Jacquemin, A, Gabszewicz, J.J. and J.A. Weymark (1983): On the stability of collusive price leadership. *Canadian Journal of Economics* 16, 17-25.

Diamantoudi, E. and E. Sartzetakis (2006): Stable international environmental agreements: An analytical approach. *Journal of Public Economic Theory* 8, 247-263.

Eichner, T. and R. Pethig (2012): Sub-global climate coalition and international trade, *CESifo working paper* 3915.

Finus, M. (2003): Stability and design of international environmental agreements: the case of transboundary pollution, in: H. Folmer and T. Tietenberg (eds.), *The International Yearbook of Environmental and Reource Economics* 2003/2004, Edward Elgar, Cheltenham, 82-158.

Hannesson, R. (2010): The coalition of the willing: Effect of country diversity in an international treaty game. *Review of International Organizations* 5, 461-474.

Finus, M. (2001): *Game Theory and International Environmental Cooperation*, Edward Elgar, Cheltenham.

Hoel, M. (1992): International environmental conventions: the case of uniform reductions of emissions *Environmental and Resource Economics* 2, 141-159.

Hoel, M. and K. Schneider (1997): Incentives to participate in an international environmental agreement. *Environmental and Resource Economics* 9, 153-170.

Kolstad, C. (2007): Systematic uncertainty in self-enforcing international environmental agreements. *Journal of Environmental Economics and Management* 53, 68-78.

Rubio, S.J. and A. Ulph (2006): Self-enforcing agreements and international trade in greenhouse emission rights. *Oxford Economic Papers* 58, 233-263.

Appendix

Appendix A: Properties of function R from (14)

Lemma 1. *The function R satisfies $\hat{s}_c := R^{-1}[(s_f = 0, m] > 0$ for all $m \in]0, n[$, $R_m(s_c, m) < 0$ for all $s_c < \hat{s}_c$, all $m \in]0, n[$, $R_{s_c s_c} = 0$ and $R_{s_c m} > 0$.*

Proof:
(i) Inserting the parametric functions (3) in

$$W^i_{e_i} = V'(e_i) + T'\left(\frac{\sum_j e_j}{n}\right) - \frac{1}{n}\left(\frac{\sum_j e_j}{n} - e_i\right) T''\left(\frac{\sum_j e_j}{n}\right) - D'\left(\sum_j e_j\right) = 0 \qquad \text{(A1)}$$

yields, after rearrangement of terms

$$e_i = \underbrace{\frac{an^2}{\alpha(2n-1) + (1+b)n^2}}_{=:G} - \underbrace{\frac{\alpha(n-1) + n^2}{\alpha(2n-1) + (1+b)n^2}}_{:=H} \sum_{j \neq i} e_j \quad \text{for} \quad i = 1, \ldots, n. \qquad \text{(A2)}$$

From (A2) we get

$$e_i = G - H\left(\sum_{j \in C} e_j + \sum_{j \in F, j \neq i} e_j\right) = G - Hme_C - H \sum_{j \in F, j \neq i} e_j \quad \text{for all}\, i \in F. \qquad \text{(A3)}$$

Summing over $i \in F$ yields

$$\sum_{i \in F} e_i = (n-m)e_f = (n-m)G - (n-m)Hme_c - (n-m-1)H(n-m)e_f \qquad \text{(A4)}$$

which can be rearranged to

$$(n-m)e_f = \frac{(n-m)G}{1 + (n-m-1)H} - \frac{(n-m)H}{1 + (n-m-1)H} me_c. \qquad \text{(A5)}$$

or equivalently to

$$s_f = R(s_c, m) = \frac{(n-m)G}{1 + (n-m-1)H} - \frac{(n-m)H}{1 + (n-m-1)H} s_c. \qquad \text{(A6)}$$

Next, verify that $\hat{s}_c := R^{-1}(s_f = 0, m) = \frac{G}{H}$ is independent of m. Finally, differentiation of (A6) yields

$$\begin{aligned} R_m &= -\frac{(1-H)G}{[1 + (n-m-1)H]^2} + \frac{(1-H)H}{[1 + (n-m-1)H]^2} s_c = -\frac{(1-H)R(s_c, m)}{(n-m)[1 + (n-m-1)H]}, \\ R_{s_c} &= -\frac{(n-m)H}{1 + (n-m-1)H} < 0, \quad R_{s_c s_c} = 0, \quad R_{s_c m} = \frac{H(1-H)}{[1 + (n-m-1)H]^2} > 0 \end{aligned} \qquad \text{(A7)}$$

due to $G > 0$ and $H \in [0, 1]$. ■

Appendix B: Properties of the functions W^c and W^f from (15) and (16) respectively

Lemma 2. *$W^c[\cdot]$ is inverse u-shaped and strictly concave in s_c, $\left(\frac{\mathrm{d}^2 W^c}{\mathrm{d}s_c^2} < 0\right)$, and $W^f[\cdot]$ is strictly decreasing in s_c, $\left(\frac{\mathrm{d}W^f}{\mathrm{d}s_c} < 0\right)$.*

Proof:

Since the coalition size m is constant throughout this proof we omit for convenience m as argument of the welfare functions. We first show the strict concavity of the coalition country's welfare function. Total differentiation of $W^c(s_c, \underbrace{R(s_c)}_{=s_f})$ from (15) yields

$$\frac{\mathrm{d}W^c}{\mathrm{d}s_c} = W^c_{s_c} + W^c_{s_f} R_{s_c}, \tag{B1}$$

$$\frac{\mathrm{d}^2 W^c}{\mathrm{d}s_c^2} = \underbrace{W^c_{s_c s_c} + W^c_{s_c s_f} R_{s_c}}_{\equiv \frac{\mathrm{d}W^c_{s_c}}{\mathrm{d}s_c}} + \underbrace{\left[W^c_{s_f s_c} + W^c_{s_f s_f} R_{s_c}\right]}_{\equiv \frac{\mathrm{d}W^c_{s_f}}{\mathrm{d}s_c}} R_{s_c} + W^s_{s_f} \underbrace{R_{s_c s_c}}_{=0}. \tag{B2}$$

Partial differentiation of

$$W^c_{s_c}(s_c, s_f) = \frac{V'\left(\frac{s_c}{m}\right)}{m} + \frac{T'\left(\frac{s_c+s_f}{n}\right)}{m} - \frac{[ms_f - (n-m)s_c]\,T''\left(\frac{s_c+s_f}{n}\right)}{n^2 m} - D'(s_c+s_f) \tag{B3}$$

yields

$$W^c_{s_c s_c} = \frac{V''}{m^2} + \frac{(2n-m)T''}{n^2 m} - \frac{[ms_f - (n-m)s_c]\,T'''}{n^3 m} - D''$$

$$= -\frac{b}{m^2} - \frac{\alpha(2n-m)}{n^2 m} - \delta, \tag{B4}$$

$$W^c_{s_c s_f} = \frac{(n-m)T''}{n^2 m} - \frac{[ms_f - (n-m)s_c]\,T'''}{n^3 m} - D'' = -\frac{\alpha(n-m)}{n^2 m} - \delta. \tag{B5}$$

Making use (B4), (B5) and $R_{s_c} = -\underbrace{\frac{(n-m)H}{(1-H)+(n-m)H}}_{=:\tilde{H}}$ (which follows from differentiation of (A6)) we get

$$\frac{\mathrm{d}W^c_{s_c}}{\mathrm{d}s_c} = -\frac{b}{m^2} - \frac{\alpha(2n-m)}{n^2 m} - \delta + \left[\frac{\alpha(n-m)}{n^2 m} + \delta\right]\tilde{H}. \tag{B6}$$

Partial differentiation of

$$W^c_{s_f}(s_c, s_f) = -\frac{[ms_f - (n-m)s_c]\,T''\left(\frac{s_c+s_f}{n}\right)}{n^2 m} - D'(s_c+s_f). \tag{B7}$$

yields

$$W^c_{s_f s_c} = \frac{(n-m)T''}{n^2 m} - \frac{[ms_f - (n-m)s_c]\,T'''}{n^3 m} - D'' = -\frac{(n-m)\alpha}{n^2 m} - \delta, \tag{B8}$$

$$W^c_{s_f s_f} = -\frac{T''}{n^2} - \frac{[ms_f - (n-m)s_c]\,T'''}{n^3 m} - D'' = \frac{\alpha}{n^2} - \delta. \tag{B9}$$

Making use of (B8), (B9) and $R_{s_c} = -\tilde{H}$ we obtain

$$\frac{\mathrm{d}W^c_{s_f}}{\mathrm{d}s_c} = -\frac{(n-m)\alpha}{n^2 m} - \delta - \left(\frac{\alpha}{n^2} - \delta\right)\tilde{H}. \tag{B10}$$

Finally, inserting (B6) and (B10) in (B2) establishes

$$\begin{aligned}\frac{\mathrm{d}^2 W^c}{\mathrm{d}s_c^2} &= -\frac{b}{m^2} - \frac{\alpha(2n-m)}{n^2 m} - \delta + \left[\frac{\alpha(n-m)}{n^2 m} + \delta\right] 2\tilde{H} + \left(\frac{\alpha}{n^2} - \delta\right)\tilde{H}^2 \\ &= -\frac{b}{m^2} - \frac{\alpha(1-\tilde{H})[2n-(1-\tilde{H})m]}{n^2 m} - \delta(1-\tilde{H})^2 \end{aligned} \tag{B11}$$

which is negative due to $\tilde{H} \in]0,1[$.

Next, we prove the monotonicity property of the fringe country's welfare function. Differentiation of $W^f(s_c, \underbrace{R(s_c)}_{=s_f})$ from (16) yields

$$\frac{\mathrm{d}W^f}{\mathrm{d}s_c} = W^f_{s_c} + W^f_{s_f} R_{s_c}, \tag{B12}$$

where

$$W^f_{s_c} = \frac{[ms_f - (n-m)s_c]\,T''}{n^2(n-m)} - D' \tag{B13}$$

$$W^f_{s_f} = \frac{V'}{n-m} + \frac{T'}{n-m} + \frac{[ms_f - (n-m)s_c]\,T''}{n^2(n-m)} - D'. \tag{B14}$$

Taking advantage of the fringe countries first-order condition (A1) which is equivalent to

$$V' + T' + \frac{[ms_f - (n-m)s_c]\,T''}{n^2(n-m)} - D' = 0 \tag{B15}$$

in (B13) and (B14) we obtain

$$W^f_{s_c} = -(V' + T'), \tag{B16}$$

$$W^f_{s_f} = -\frac{n-m-1}{n-m}(V' + T'). \tag{B17}$$

Inserting (B16) and (B17) in (B12) we get

$$\frac{\mathrm{d}W^f}{\mathrm{d}s_c} = -(V' + T')\left[1 + \frac{n-m-1}{n-m} R_{s_c}\right]. \tag{B18}$$

Since the terms in square brackets are positive $\frac{\mathrm{d}W^f}{\mathrm{d}s_c} < 0$ holds, if and only if $V' + T' > 0$. From (5) and $V'(e_f) = p + \pi_f$ (which follows from the fringe countries' consumers utility maximization) we have $V' + T' = \pi_f$. From (A1) we infer that $V' + T' > 0$ if $e_f > e_c$. Finally, it can be shown that π_f remains positive when the coalition relaxes its emission cap and the fringe countries tighten their emission caps. ■

Appendix C: Proof of Proposition 1

Account for $\frac{\mathrm{d}(s_c+s_f)}{\mathrm{d}s_c} = 1+R_{s_c}$, and determine the first-order condition for an interior solution to (17),

$$\begin{aligned}\frac{\mathrm{d}(mW^c)}{\mathrm{d}s_c} &= W^c_{s_c} + W^c_{s_f}R_{s_c} \\ &= V' + T' - \left(\frac{s_c+s_f}{n} - \frac{s_c}{m}\right)\frac{m(1+R_{s_c})T''}{n} - m(1+R_{s_c})D' = 0. \quad \text{(C1)}\end{aligned}$$

If the coalition of any size $m \in [1,n[$ chooses the strategy $s_c = me_o$, the fringe's best reply is $s_f = R(me_o, m) = (n-m)e_o$ and the BAU equilibrium results. At that equilibrium, i.e. evaluated at $s_c = me_o$, the coalition's marginal welfare is

$$\begin{aligned}MWC_o(m) &:= \frac{\mathrm{d}(mW^c)}{\mathrm{d}s_c}\Big|_{s_c=me_o} = \\ &= \underbrace{V'(e_o)+T'(e_o)}_{\text{marginal consumption welfare, same for all coaltion sizes}} + \underbrace{\left\{-D'(ne_o) + [1-m(1+R_{s_c}]D'(ne_o)\right\}}_{\text{marginal climate welfare for } m\in]1,n[}. \quad \text{(C2)}\end{aligned}$$

According to (C2) the coalition's marginal consumption welfare is independent of m, while its marginal climate welfare is not. Since by definition of $\tilde{m}$ the condition $\tilde{m}[1+R_{s_c}(\tilde{m}e_o, \tilde{m})] = 1$ is satisfied, the equations (C1) and (C2) yield for a coalition of size

$$MWC_o(\tilde{m}) = \underbrace{V'(e_o)+T'(e_o)}_{\text{marginal consumption welfare}} + \underbrace{[-D'(ne_o)]}_{\text{marginal climate welfare for } m=\tilde{m}} = 0. \quad \text{(C3)}$$

(C3) is identical to the first-order condition of all n countries in the non-cooperative BAU scenario of Section 2. We invoke (C3) to rewrite (C2) as

$$\begin{aligned}MWC_o(m) &= \underbrace{V'(e_o)+T'(e_o)-D'(ne_o)}_{=0} + [1-m(1+R_{s_c})]D'(ne_o) \\ &= [1-m(1+R_{s_c})]D'(ne_o). \quad \text{(C4)}\end{aligned}$$

(C4) holds for any given $m \in [1,n[$. Since $\frac{\mathrm{d}[m(1+R_{s_c})]}{\mathrm{d}m} = (1+R_{s_c}) + mR_{s_cm} > 0$, the equivalence $\left\{[1-m(1+R_{s_c})] \gtreqless 0 \iff m \lesseqgtr \tilde{m}\right\}$ holds. Finally, differentiation of (A6) with respect to s_c yields $R_{s_c} = -\frac{(n-m)H}{1+(n-m-1)H}$. Inserting this term in $[1-\tilde{m}(1+R_{s_c})] = 0$ we get $\tilde{m} = 1+(n-1)H$. Making use of the definition of H from (A2) establishes (19) after some rearrangement of terms. ■

Appendix D: Proof of (23)

If $m < \tilde{m}$, (21) implies $m\mathcal{E}^c(m) > me_o$ and therefore $\mathcal{W}^f(m) < W_o$ because $\frac{\mathrm{d}W^f}{\mathrm{d}s_c} < 0$ due to Appendix B. If $m > \tilde{m}$, (21) implies $m\mathcal{E}^c(m) < me_o$ and therefore $\mathcal{W}^f(m) > W_o$ because $\frac{\mathrm{d}W^f}{\mathrm{d}s_c} < 0$ due to Appendix B. Analogously, $\mathcal{W}^f(m) > \mathcal{W}^c(m)$ for $m > \tilde{m}$ follows from $m\mathcal{E}(m) < me_o$ and Appendix B.

Appendix E: Proof of Proposition 4

By construction, the functions Ω^c and Ω^f approximate the functions $\mathcal{W}^c$ and $\mathcal{W}^f$ on the sub-domain $[\tilde{m}, n]$, because the properties

$\mathcal{W}^f(\tilde{m}) = \mathcal{W}^c(\tilde{m}) = W_o$ and $\mathcal{W}^f(m) > \mathcal{W}^c(m) > W_o,\ m > \tilde{m}$ (equivalence (23)),
$\mathcal{W}^f_m(\tilde{m}) > 0, \mathcal{W}^c_m(\tilde{m}) = 0$ and $\mathcal{W}^h_m(m) > 0, h = c, f, m > \tilde{m}$ (Figure 2), and
$\mathcal{W}^f_m(m) - \mathcal{W}^c_m(m) > 0$ for $m > \tilde{m}$ and increasing in m (Figure 2),

of the functions $\mathcal{W}^c$ and $\mathcal{W}^f$ are satisfied by the functions Ω^c and Ω^f.

Ad (i): For analytical convenience we consider here the function $\hat{H} : [\tilde{m}, n] \longrightarrow \mathbb{R}_+$ defined by $\Big[h = \hat{H}(m) \quad \Longleftrightarrow \quad \mathcal{W}^f(m) = \mathcal{W}^c(m+h)\Big]$ (rather than applying the function H defined in the text). We replace $\mathcal{W}^c$ and $\mathcal{W}^f$ by Ω^c and Ω^f from (28) and solve the equation $\Omega^f(m) = \Omega^c(m+h)$ and obtain

$$h = \hat{H}(m, \omega_1, \omega_2, \omega_3) = -(m - \tilde{m}) + \sqrt{\frac{2\omega_2(m-\tilde{m}) + (\omega_1+\omega_3)(m-\tilde{m})^2}{\omega_1}}. \tag{E1}$$

From (E1) we get $\hat{H}(m) = 0$ for $m = \tilde{m}$ and $\hat{H}(m) > 0$ for $m > \tilde{m}$. For all $m \geq \tilde{m}$ the first derivative is

$$\hat{H}_m = -1 + \rho > 0, \quad \text{where } \rho := \sqrt{1 + \frac{\omega_2^2 + [2\omega_2 + (\omega_1+\omega_3)(m-\tilde{m})]\omega_3(m-\tilde{m})}{[2\omega_2 + (\omega_1+\omega_3)(m-\tilde{m})]\omega_1(m-\tilde{m})}}. \tag{E2}$$

$\hat{H}(0) = 0$ and (E2) imply that there is one and only one $\bar{m} \in]\tilde{m}, n[$ satisfying $\hat{H}(\bar{m}) = 1$. Hence if $\bar{m}$ is an integer and $\bar{m} > \tilde{m}$, the coalitions of size $\bar{m}$ and size $\bar{m}+1$ are stable coalitions. Otherwise, there exists one and only one stable coalition. Its size is the (unique) integer in the interval $]\bar{m}, \hat{m}+1[$.

Ad (ii): Verify $\hat{H}_{\omega_1} = -\frac{2\omega_2(m-\tilde{m})+\omega_3(m-\tilde{m})^2}{2\rho\omega_1^2} < 0$, $\hat{H}_{\omega_2} = \frac{m-\tilde{m}}{\rho\omega_1} > 0$, $\hat{H}_{\omega_3} = \frac{(m-\tilde{m})^2}{2\rho\omega_1} > 0$ and observe that the differential of $\hat{H}(\bar{m}, \omega_1, \omega_2, \omega_2) = 1$ yields $\frac{\partial\bar{m}}{\partial\omega_i} = -\frac{\hat{H}_{\omega_i}}{\hat{H}_m}$ for $i = 1, 2, 3$. Therefore sign $\frac{\partial\bar{m}}{\partial\omega_i} = -\text{sign}\, \hat{H}_{\omega_i}$.

Ad (iii): Since $\hat{H}_m > 0$, $\bar{m} \in]\tilde{m}+1, \tilde{m}+2[$ implies $\hat{H}(\tilde{m}+2) > 1$. Solving (E1) for $m = \tilde{m}+2$ yields $\hat{H}(\tilde{m}+1) = -1 + \sqrt{\frac{\omega_1+2\omega_2+\omega_3}{\omega_1}}$ and hence $\hat{H}(\tilde{m}+1) > 1 \quad \Longleftrightarrow \quad 3\omega_1 < 2\omega_2 + \omega_3$. From $\hat{H}_m > 0$ and $\hat{H}(\tilde{m}) = 0$ follows that $\hat{H}(\tilde{m}+1) > 1$ implies $\bar{m} \in]\tilde{m}, \tilde{m}+1[$. ■

Appendix F: Proof of $\left|R_{s_c}\right| > \left|R^a_{s_c}\right|$

Making use of the parametric functions in the fringe country's first-order condition $V'(e_i) + T'(e_i) - D'\left(\sum_j e_j\right) = 0$ yields

$$e_i = \frac{\alpha}{\alpha+b+1} - \frac{1}{\alpha+b+1}\sum_{j\neq i} e_j. \tag{F1}$$

Multiplying (F1) by $(n-m)$ and setting $e_i = e_f = \frac{s_f}{n-m}$ and $\sum_{j\neq i} e_j = me_c + (n-m-1)e_f = s_c + \frac{n-m-1}{n-m}s_f$ we obtain after rearrangement of terms the aggregate fringe best reply function

$$s_f = R^a(s_c, m) := \frac{(n-m)\alpha}{\alpha+b+n-m} - \frac{n-m}{\alpha+b+n-m}s_c. \tag{F2}$$

Next, differentiating (A6) and (F2) we get

$$\left|R^a_{s_c}\right| < \left|R^a_{s_c}\right| \iff \frac{1}{\alpha+b+n-m} < \frac{H}{1+(n-m-1)H} \iff \frac{1}{H} < 1+\alpha+b. \tag{F3}$$

Inserting H from (A2) in (F3) and rearranging terms establishes

$$\left|R^a_{s_c}\right| < \left|R^a_{s_c}\right| \iff \alpha n < \alpha(\alpha+b)(n-1) + \alpha n^2. \tag{F4}$$

■

Appendix G: Proof of Proposition 5(ii)

We show that $m^*_a - \tilde{m}_a \leq 2$ for all $(\check{b}, n) \in \Lambda := \{(\check{b}, n)\big|\check{b} > \frac{n(n-4)}{4}, n > 4\}$ by inserting $\check{b} = n(n-4)/4$ in (32) and making use of $\frac{\mathrm{d}\tilde{m}_a}{\mathrm{d}b} < 0$ to obtain[33]

$$\tilde{m}_a \in]1, \bar{M}^a(n)[\quad \text{where } \bar{M}^a(n) := \frac{n^2}{n^2-4(n-1)}. \tag{F5}$$

Closer inspection of (F5) reveals that $\bar{M}^a(5) = 2.77$ and that $\frac{\mathrm{d}\bar{M}^a(n)}{\mathrm{d}n} < 0$ for $n > 4$. Hence we get

$$\tilde{m}_a \in]1, 2.77[\quad \text{for all } (\check{b}, n) \in \Lambda. \tag{F6}$$

In view of (32) and (F6) and $m^*_a \leq 3$ we conclude that $m^*_a - \tilde{m}_a \leq 2$ for all $n > 4$.

[33]The observation that 1 is a lower bound for $\tilde{m}_a$ follows directly from (32).