Conditional pledges in climate agreement negotiations: How far can they carry?

Draft: Do not cite or circulate.

Florian Landis*

April 17, 2016

Abstract

In negotiating for a global climate agreement, the EU has previously made its contribution to GHG emissions reductions conditional on other countries' contributions. This paper analyzes whether this conditionality can motivate other countries to increase their contributions and whether it gives them an incentive to join the EU in a coalition that makes its efforts conditional on other countries' actions. My results suggest that the EU's strategy can only induce meaningful additional emission reductions in big non-coalition countries. Similarly, if joining the coalition is to be attractive for a non-coalition country, that country has to be of at least similar size as the existing coalition.

Keywords: climate change negotiations; conditional abatement; coalition stability JEL: F53; H87; Q54;

*ETH Zürich
Chair for Economics/Energy Economics
ZUE E3
Zürichbergstrasse 18
8092 Zürich
Switzerland
E-mail: landisf@ethz.ch
Tel: +41 44 633 8453

1 Introduction

At the negotiating table for a global climate agreement, the European Union (EU) has committed itself to emission targets unconditionally, while at the same time promising to increase the ambition of the targets, if other nations follow the EU in making meaningful contributions to the effort of globally reducing carbon emissions. There are at least three ways how this negotiation strategy can be motivated. Firstly, it has been observed that individual contributions to a public good may be positively correlated to the contributions of others. This correlation has been observed without evidence of any individual's contributions having an effect on either cost or benefits of contributing for any other individual. For example, Frey and Meier (2004) report that students' contributions to charitable funds depend on what they are told about other students' propensity to donate. In the case of the EU, the contributions of its counterparts are uncertain. Thus, the EU may have chosen to make its own contribution conditional on the others' in order to hedge against this uncertainty.

Secondly, any given level of abatement effort by the EU becomes less costly as the remaining countries abate more of their emissions. This cost reduction is brought about by gained competitiveness of European firms in a more level playing field created by increased regulation abroad. As increased emission reductions by the EU have the same effect on costs in the remaining countries, the promise by the EU to increase abatement if a counterpart does likewise will actually reduce the expected cost of doing so for that counterpart. This line of argument is discussed by Underdal et al. (2012) as a motive for making conditional contributions. Their work analyzes the EU's conditional contributions and argues that it would be politically realistic for the EU to honor its promises. However, their assessment of the likelihood that other important countries will react to the EU's conditional contribution by adopting a more stringent policy is mixed.

The third argument is based on the observation that the conditionality of

the offer changes selfish incentives of the EU's counterparts for setting their contribution levels. If they believe the EU's offer, a raise in their abatement efforts promises to create the additional benefits of the EU's additional abatement as well. Ignoring the interdependence of abatement cost discussed by Underdal et al., this paper sets out to analyze to what extent such a negotiating strategy can increase the abatement efforts of other world regions. As a side, I ask if the strategy gives an incentive for other countries to join Europe in a matching coalition.

While conceptually promising, this idea of matching other countries' emission reductions by increasing one's own reduction efforts can be applied in a more rigorous way than the EU has been doing: the EU could refrain from making unconditional promises and make all action against climate change conditional on global abatement efforts. It is this extreme form of matching that I consider in this paper. The results of my analysis will then provide an upper bound on what a coalition faced by selfishly acting counterparts can potentially achieve by making conditional contributions to the abatement effort.

In order to make matching rates credible, the matching coalition has to have near-indifference between a wide range of abatement effort levels or preference for rising abatement levels if other countries increase their efforts. On the one hand, it is hard to decide if the EU is united enough politically for each country to internalize all externalities on other countries, or if each country acts out of its own self-interest when it sets its policies. On the other hand, studies of global climate policies suggest that by price, trade, and competitiveness effects, the cost for ambitious EU climate policy decreases as the global effort level increases (this was the effect exploited in the discussion of Underdal et al. (2012)). Both these observations lend some credibility to the notion that the EU should indeed see it as incentive compatible to follow different abatement plans dependent on abatement levels reached by the non-EU world.

Additionally to analyzing how much such conditional contributions can increase abatement outside the EU, I consider the incentives for outsider countries

to form a coalition with the matching countries. My findings indicate that, due to free-rider incentives, introducing matching does not facilitate the expansion of small coalitions into stable bigger ones. Thus, the strategy of conditional contribution will fail to lead the world to Pareto-efficient levels of emission reductions. At least, if negotiating countries are big and thus internalize a big part of the global climate externality, matching can play a meaningful role even if only few countries join the coalition.

In the following pages I give a mathematical formulation of the matching and emission reduction decisions taken by coalition and non-coalition countries (Section 2). I then illustrate the concept of matching and coalition formation with the example of linear marginal abatement cost curves and identical world regions (Appendix A). The theoretical results indicate that, as the size of political units that negotiate the provision of the public good decreases, big coalitions are not stable even if they can use matching and thus fail to provide a meaningful amount of the public good. I additionally calibrate the stylized model to results for cost and benefits from emission reductions from the RICE integrated assessment model (Nordhaus and Yang, 1996; Nordhaus, 2010) (Section 3). The results in this context indicate that the asymmetry of costs and benefits across the twelve regions of the model make a coalition possible that would not be stable if all regions were the same. On the other hand, many coalitions that would be stable in a symmetrical setting become unstable in the asymmetrical context. Section 4 sets my results into context, discusses the usefulness of matching in negotiations over global climate policy and names some caveats of my analysis.

2 Model

For the purpose of describing the model, I assume the world to be divided into n political units ('countries'), k of which are assumed to build a coalition. The countries within the coalition are indexed by c, while those outside the coalition are indexed by nc. The countries face different but constant marginal benefits b_c

and b_{nc} from global emission reductions and marginal abatement costs $mc_c(a_c)$ and $mc_{nc}(a_{nc})$ for locally reducing emissions.

Faced with the coalition's matching rate μ, the non-coalition countries abate until their private marginal benefits from their own and matched emission reductions equal the marginal cost

$$\mathrm{mc}_{nc}(a_{nc}) = (1+\mu)\mathrm{mb}_{nc}. \tag{1}$$

The coalitions decisions, however, are more involved. The coalition wants to abate at a level at which benefits and costs of abating balance each other. But at the same time, it does not want to achieve this level of abatement unconditionally, but it tries to reach it by matching the action of the non-coalition countries. The non-coalition abatement levels $\hat{a}_{nc}$ without matching are determined by

$$\mathrm{mc}_{nc}(\hat{a}_{nc}) = \mathrm{mb}_{nc},$$

and thus are independent of other countries' abatement decisions. The coalition can credibly threaten to reduce abatement to levels $\hat{a}_c$ itself. Thus, the matching rate it wants to announce is such that its abatement beyond $\hat{a}_C := \sum_c \hat{a}_c$ is the announced matching rate μ times the non-coalition countries' abatement beyond $\hat{a}_{nc}$:

$$a_C - \hat{a}_C = \mu \sum_{nc} \left(a_{nc} - \hat{a}_{nc}\right). \tag{2}$$

When the coalition chooses abatement levels a_c, it not only expects benefits from these abatement levels, but it also anticipates an increase in the matching rate that it can provide and thus an increase in non-coalition abatement. It therefore expects infinitesimal increases $\mathrm{d}a_C$ in abatement to result in increases $\mathrm{d}\mu$ of the matching rate and $\mathrm{d}a_{nc}$ of non-coalition abatement. Equation (2)

implies

$$\begin{aligned} \mathrm{d}a_C &= \mathrm{d}\mu \sum_{nc} (a_{nc} - \hat{a}_{nc}) + \mu \sum_{nc} \mathrm{d}a_{nc} \\ &= \mathrm{d}\mu \left[\sum_{nc} (a_{nc} - \hat{a}_{nc}) + \mu \underbrace{\sum_{nc} \frac{\partial a_{nc}}{\partial \mu}}_{=:\alpha} \right]. \end{aligned}$$

According to equation (1), a_{nc} is only influenced by μ and

$$\frac{\partial \mathrm{mc}_{nc}}{\partial a_{nc}} \mathrm{d}a_{nc} = \mathrm{d}\mu \mathrm{mb}_{nc}.$$

It follows that

$$\frac{\partial a_{nc}}{\partial \mu} = \frac{\mathrm{mb}_{nc}}{\frac{\partial \mathrm{mc}_{nc}}{\partial a_{nc}}},$$

which gives α as

$$\alpha = \sum_{nc} \frac{\mathrm{mb}_{nc}}{\frac{\partial \mathrm{mc}_{nc}}{\partial a_{nc}}}.$$

Summarizing, I obtain

$$\begin{aligned} \mathrm{d}\left(\sum_{nc} a_{nc} \right) &= \sum_{nc} \frac{\partial a_{nc}}{\partial \mu} \cdot \frac{\partial \mu}{\partial a_C} \mathrm{d}a_C \\ &= \frac{\alpha}{\sum_{nc} (a_{nc} - \hat{a}_{nc}) + \mu\alpha} \mathrm{d}a_C, \end{aligned}$$

and thus for every infinitesimal increase $\mathrm{d}a_C$ of coalition abatement, global abatement increases by $\left(1 + \frac{\alpha}{\sum_{nc}(a_{nc} - \hat{a}_{nc}) + \mu\alpha}\right) \mathrm{d}a_C$. The first-order conditions for setting the coalition countries' abatement levels are therefore

$$\mathrm{mc}_c(a_c) = \left(1 + \frac{\alpha}{\sum_{nc} (a_{nc} - \hat{a}_{nc}) + \mu\alpha} \right) \sum_{c'} \mathrm{mb}_{c'}. \tag{3}$$

Equations (1)–(3) determine equilibrium abatement and the matching behavior if the coalition countries internalize benefits across the coalition but abate

accordingly only if non-coalition countries increase their abatement. This conditionality provides the incentive for the additional non-coalition abatement in the first place.

3 The RICE setting

In the following I introduce realistic representations of cost and benefits of climate change for different countries and world regions. In a first step, I evaluate the model predictions using the Regional Integrated Climate-Economy model (RICE) (Nordhaus and Yang, 1996; Nordhaus, 2010) with features 12 world regions and represents them with individual cost curves for emission reductions and damage functions from future climate warming. This allows me to study the effect of asymmetry between regions on the outcome of conditional abatement by a coalition. I look at the stability of coalitions of different size and find that only small coalitions are stable; they are not able to attract a significant number of outsider regions to join the coalition. In a second step, I disaggregate the world regions into single countries. This is important for estimating the effect of conditional abatement on non-coalition countries' behavior because aggregated regions internalize unrealistically high shares of global damages and thus react too strongly to matching rates. I posit that the EU represents a 28-country coalition that does not have to worry about internal stability of the coalition. In this setting, I estimate by how much the EU's negotiation strategy can increase global abatement and analyze if additional (large) countries would prefer joining the coalition to staying outside.

The RICE (Nordhaus and Yang, 1996; Nordhaus, 2010) has been built for an analyzing the trade-offs of emission reductions in the context of climate change. It describes abatement cost and vulnerability to climate change for 12 regions: China, USA, EU, Mid-East, India, Other Asia, Russia, Latin America, Japan, Eurasia, Africa, and Other High Income Countries (OHI). For the period labeled

2015, the model predicts the cost of abatement for region r to be

$$c_r(a_r) = \frac{p_{bs,r} y_r \sigma_r}{2.8} \left(\frac{a_r}{y_r \sigma_r} \right)^{2.8},$$

where $p_{bs,r}$ denotes the regional price level of backstop technologies, y_r is regional gross domestic product (GDP) before abatement efforts, and σ_r is the emission intensity of GDP. Marginal abatement costs then are

$$\mathrm{mc}_r(a_r) = \underbrace{\frac{p_{bs,r}}{(y_r \sigma_r)^{1.8}}}_{\beta_r} a_r^{1.8}.$$

The discounted benefits of abatement (assuming socially optimal abatement à la Nordhaus in consecutive periods) are also regional and are labeled mb_r.[1] These region-specific characterizations of costs and benefits from emission abatement allow the analysis of how a realistic degree of asymmetry can influence the possible formation of stable coalitions in the setting of matching coalitions.

The abatement levels without coordinated climate policy in 2015 are

$$\hat{a}_r = \left(\frac{\mathrm{mb}_r}{\beta_r} \right)^{1/1.8} = y_r \sigma_r \left(\frac{\mathrm{mb}_r}{p_{bs,r}} \right)^{1/1.8}$$

and an infinitesimal increase in the matching rate of the coalition increases non-coalition abatement by

$$\begin{aligned} \alpha &= \sum_{nc} \frac{\mathrm{mb}_{nc}}{\frac{\partial \mathrm{mc}_{nc}}{\partial a_{nc}}} \\ &= \sum_{nc} \frac{\mathrm{mb}_{nc}}{1.8 \beta_{nc} a_{nc}^{0.8}}. \end{aligned}$$

The first-order conditions solved by regions inside and outside the coalition

[1] The marginal benefits from abatement in the RICE model are by no means constant by design. But numerical evaluation of the marginal benefits of global abatement in a single period shows that they hardly depend on the level of global abatement.

therefore are

$$\beta_{nc} a_{nc}^{1.8} = (1+\mu)\text{mb}_{nc} \tag{4}$$

$$\sum_{c} (a_c - \hat{a}_c) = \mu \sum_{nc} (a_{nc} - \hat{a}_{nc}) \tag{5}$$

$$\beta_c a_c^{1.8} = \left(1 + \frac{\sum_{nc} \frac{\text{mb}_{nc}}{1.8\beta_{nc} a_{nc}^{0.8}}}{\sum_{nc}(a_{nc} - \hat{a}_{nc}) + \mu \sum_{nc} \frac{\text{mb}_{nc}}{1.8\beta_{nc} a_{nc}^{0.8}}}\right) \sum_{c'} \text{mb}_{c'}. \tag{6}$$

Given the asymmetric costs and benefits from abatement of different regions, numerous different coalitions can be built. For all possible coalition, I compute abatement levels and compare the welfare levels that the different coalition and non-coalition regions achieve under the given configurations. The results from such comparisons tell if a given coalition is stable (no region wants to leave the coalition) or if it is expandable (at least one non-coalition region would like to join the coalition and no coalition region objects). Table 1 shows aggregate statistics for stability and expandability of coalitions of a given size.

Table 1: **Stability and expandability.** Columns three to eight show how many of the total possible numbers of coalitions (column two) of a given size (column one) are stable or expandable. A coalition is termed stable if no single coalition member wants to leave the coalition. A coalition is termed expandable if a non-coalition region wants to join the coalition and no coalition member objects this. Coalitions' behavior is differentiated between only taking their own abatement (OA) into account when deciding on abatement levels and between incorporating the fact that increasing their own abatement gives increased non-coalition abatement through an increased matching rate (OMA).

Size	#	Stability			Expandability		
		no matching	OA	OMA	no matching	OA	OMA
1	12	12	12	12	11	12	12
2	66	14	53	23	23	64	45
3	220		63	1	21	137	50
4	495		10		13	147	22
5	792				2	89	3
6	924					20	
7	792						
8	495						
9	220						
10	66						
11	12						
12	1						

The results indicate that matching indeed increases both internal stability and expandability of coalitions vis-à-vis the no-matching case. It is remarkable that if the coalition ignores the fact that increases in its own abatement lead to increases in non-coalition abatement (replace the parentheses in equation (6) by 1; I call this the OA setting), this encourages stability further. As including non-coalition abatement in its decision (here called the OMA setting) increases the welfare of the coalition, it makes being in the coalition more attractive. But at the same time, by increasing global abatement, it also makes free-riding in the non-coalition more attractive, and this appears to be the more important effect.

In order to assess the effect of the asymmetry between regions on the formation of coalition, I compare the results for stability in the context of RICE with the results for the symmetrical case. For this, I aggregate global cost of global abatement, add up global benefits[2] and redistribute them symmetrically between 12 regions. Figure 1 shows how welfare for coalition and non-coalition regions compares for different coalition sizes. In the OMA setting, coalitions of

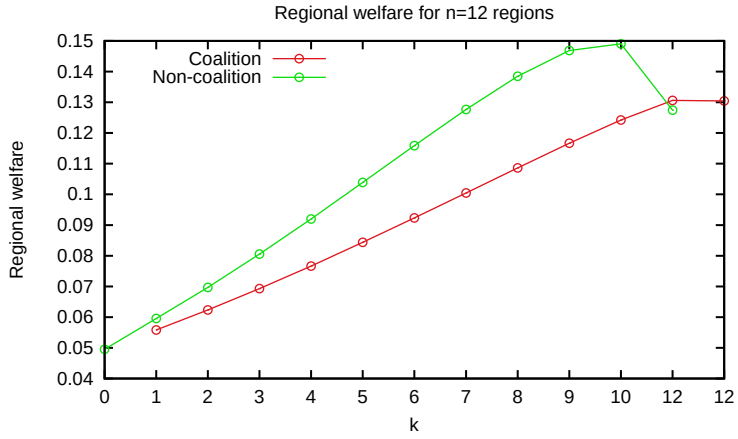

Figure 1: Welfare for regions inside and outside a given matching coalition of size k for global abatement cost according to the RICE model. Only coalitions of size $k = 1, 2$ are internally stable.

[2]The global abatement cost function in RICE is $0.0276 \cdot A^{2.8}$ and global marginal benefits are 0.384. Normalizing global marginal benefits to 1 gives $\gamma = 2.8$ and $\beta = 2.8 \cdot 0.0276/0.384 = 0.201$ (compare to Appendix B).

size 2 are stable, but larger coalitions are not, and in the OA setting coalitions of size up to 4 are stable. Comparing this with the asymmetrical OMA setting shows that in the asymmetrical case, only 23 out of 66 coalitions of size 2 are stable, whereas in the symmetrical case, coalitions of 2 regions are always stable. However, the asymmetrical case offers the possibility of forming one coalition with 3 member regions, something that would not be possible in the symmetrical setting. In the OA setting, coalitions of size 4 that are always stable in the symmetrical setting are only stable in 10 out of almost 500 cases. And as in the symmetrical setting, the asymmetrical setting does not allow internally stable coalitions of 5 regions.

The value to the global community of the different coalitions in an asymmetrical world should be assessed by the abatement these coalitions are able to generate. I abstain from doing this in the setting of the RICE model, because my analysis in Appendix A has shown that abatement levels in the presence of a coalition depend on the size of the regions in the model. As the regions of the RICE model do not reflect the actual political units on planet earth, region size is inaccurately represented and the abatement results would be wrong. I cannot even realistically analyze the case of one large country like China, India, Russia, or the USA joining a coalition with the EU, because my model in its current form misrepresents the EU's threat point, in which all member states only internalize the benefits from abatement on their own territory rather than the benefits for the whole EU. Thus, the current analysis within the RICE model only serves to analyze the effect that asymmetry has on the stability and expandability of coalitions.

3.1 Subdividing the RICE regions

The representation of twelve regions using benefits and cost of abatement according to the RICE model introduces a realistic amount of asymmetry to the setting. But by misrepresenting the number and size of actual countries, the above setting is unable to answer the question of how many countries the EU

would be able to attract into a coalition and by how much non-coalition abatement can be increased. This is due to the fact that the EU is represented by one single region and in reality has a lower threat point for designing its matching scheme, as the single countries of the Union may threaten to only internalize own benefits of abatement if the rest of the world does not abate more than is in their self interest. The higher matching rates decrease the attractiveness of staying out of the coalition for large regions like China, the US, India, Russia, and Japan. On the other hand, we've seen that small regions who initially internalize a small share of global abatement, profit less from joining a large coalition which would require them to increase abatement considerably. Thus the actual countries within composite regions like Mid-East, Other Asia, Latin America, Eurasia, and OHI, might be much less inclined to join the coalition than what above simulations based on the composite regions may indicate.

To resolve these questions, I split the RICE regions into single countries. I make the somewhat heroic assumption that within RICE regions, countries have the same characteristics (emission intensity of the economy, exposure to climate damages, abatement options) but differ only in size. I use emissions data from the World Resource Institute[3] to determine the 'size' of the countries. I assume that the EU constitutes a coalition from which defection is not an issue (costs and benefits of emissions abatement only make up a part of the benefits of staying in the EU, after all). The remaining question is, if the the EU can bring one or two other large countries to join a coalition using the matching mechanism. Carrying out the numerical analysis, I find that the three large countries U.S., China, and India would indeed join the coalition. Further extension of the coalition is not possible however. Even though both India and China would be willing to join a EU+US coalition, the US would want prefer to leave the resulting EU+US+X coalition. Only India would like to join the EU+China coalition, but again, China would want to leave the resulting coalition. No further country would like to join a EU+India coalition. I conclude that in

[3] `cait.wri.org`

terms of coalition extension, the matching strategy would have the theoretical potential of bringing an additional large emitter into the coalition.

In terms of abatement potential of this strategy, Table 2 shows that by implementing the matching mechanism alone, the EU could increase global abatement by about 0.10 GtC.[4] By building a larger coalition together with China or the United States, another 0.14 GtC or 0.11 GtC could be abated. Thus the EU would have the opportunity to more than double the effectiveness of the coalition-matching by getting another large country to join the coalition, which would be in the self-interest of that country.

Coalition	Abatement	Abatement w/o matching
EU	0.478	0.374
EU+China	0.617	
EU+US	0.591	
EU+India	0.553	

Table 2: Global abatement of different stable coalitions

4 Conclusion

In this paper, I analyze the strategic benefits of making ones contribution to emissions reductions in the negotiations about a global climate agreement conditional on other countries' contributions. I find that in a situation where large non-identical countries negotiate with each other, this strategy can have important benefits. Unilateral matching, as this strategy may be called, can both increase the contribution of the matched countries and encourage them to join a matching coalition up to a certain size. If the matched countries are small, however, the fact that they each internalize a smaller share of global benefits from abatement makes the matching mechanism less effective.

The fact that, even under such favorable modeling choices, matching seems of limited use for significantly increasing global abatement levels makes match-

[4] In 2015, annual baseline emissions of industrial GHGs are 10.0 GtC according to the RICE model and optimal abatement would be 1.5 GtC.

ing an unlikely candidate to bring a break-through in the current stalemate of international climate negotiations. What is left is the idea that such conditionality of abatement efforts could be a model for other international unions and thus eventually lead to a full reciprocal matching scheme à la Guttman and Schnytzer (1992), which could lead to Pareto-efficient global outcomes, as a well established theoretical literature suggests.

My analysis makes several assumptions that are favorable for a meaningful effect of conditional commitments to emission abatement. For one, the coalition fully exploits the concept of conditional contributions to its advantage. However, this might not be a realistic course of action in a negotiation setting ruled by diplomacy. Also, in order for the mechanism to be effective, the non-coalition countries must react rationally to conditional offers without taking into account their own negotiation strategies or the argument about fair distribution of abatement cost that are so important in those negotiations. Even so, my analysis finds little hope for a real breakthrough coming from a single coalition that offers conditional contribution levels to global abatement.

But as mentioned in the introductory remarks, the interdependency of regional abatement costs also plays a role in favor of announcing conditional contributions to emission abatement. Additionally, such conditional contributions may also invoke a desire for reciprocity and induce additional abatement without the non-coalition drawing direct material benefit from that. Thus, announcing conditional contributions when entering into negotiations may still be a strategy that is worthwhile pursuing if promises of additional contributions are both meaningful and credible.

References

Bruno S Frey and Stephan Meier. Social comparisons and pro-social behavior: Testing "conditional cooperation" in a field experiment. *American Economic Review*, 94(5):1717–1722, December 2004.

Joel M. Guttman and Adi Schnytzer. A solution of the externality problem using strategic matching. *Social Choice and Welfare*, 9(1):73–88, January 1992.

William D. Nordhaus. Economic aspects of global warming in a post-copenhagen environment. *Proceedings of the National Academy of Sciences*, 107(26):11721–11726, June 2010.

William D. Nordhaus and Zili Yang. A regional dynamic general-equilibrium model of alternative climate-change strategies. *The American Economic Review*, 86(4):741–765, September 1996.

A. Underdal, J. Hovi, S. Kallbekken, and T. Skodvin. Can conditional commitments break the climate change negotiations deadlock? *International Political Science Review*, 33(4):475–493, September 2012.

A The symmetrical situation with quadratic cost function

Assume n countries, and assume k of them to form a coalition. All countries have same constant marginal benefit $\mathrm{mb} = 1/n$ from abatement and face an abatement cost curve $\mathrm{c}(a) = n\beta a^2/2$, making their marginal abatement cost $\mathrm{mc}(a) = n\beta a$.[5]

It would be socially optimal if all countries were to abate $1/(n\beta)$ units of emissions, as

$$\sum_{j=1}^{n} \mathrm{mb}_j = 1 = \mathrm{mc}_i\left(\frac{1}{n\beta}\right) \qquad \forall i.$$

If all countries only take their own benefits into account, abatement is

[5]Global abatement $A = na$, then causes global abatement cost $C(A) = n \cdot c(A/n) = \beta A^2/2$. Thus global abatement cost is independent of n and n can be interpreted as the number of regions into which a given world is divided.

$1/(n^2\beta)$:

$$\mathrm{mb}_i = 1/n = \mathrm{mc}_i\left(\frac{1}{n^2\beta}\right) \qquad \forall i.$$

Given the overall matching rate of the coalition μ, the first-order conditions for the coalition c and non-coalition countries nc are

$$\begin{aligned} n\beta a_c &= \left(1 + \frac{\alpha}{(n-k)\left(a_{nc} - \frac{1}{n^2\beta}\right) + \mu\alpha}\right)\frac{k}{n} \\ n\beta a_{nc} &= \frac{(1+\mu)}{n}. \end{aligned}$$

In agreement with the Section 2, α is determined by

$$\begin{aligned} n\beta \mathrm{d}a_{nc} &= \frac{\mathrm{d}\mu}{n} \\ \Rightarrow \alpha = (n-k)\frac{\partial a_{nc}}{\partial \mu} &= \frac{(n-k)}{n^2\beta}. \end{aligned}$$

Thus, the first-order conditions become

$$n^2\beta a_c = k\left(1 + \frac{1}{n^2\beta a_{nc} - 1 + \mu}\right) \tag{7}$$

$$n^2\beta a_{nc} = 1 + \mu. \tag{8}$$

Full use of a_c for matching (the threat point is the non-coordination level of abatement) requires

$$\begin{aligned} k\left(a_c - \frac{1}{n^2\beta}\right) &= \mu(n-k)\left(a_{nc} - \frac{1}{n^2\beta}\right) \\ \Rightarrow \mu &= \frac{k(n^2\beta a_c - 1)}{(n-k)(n^2\beta a_{nc} - 1)} \end{aligned} \tag{9}$$

Appendix C shows that the solution to equations (7)–(9) exists and is unique.

A.1 Asymptotic properties

The global benefits and total costs of global abatement remain constant independently of the number of regions n into which the world is divided. Varying n but keeping k proportional reveals how the coalition-cum-matching strategy performs for different levels of regional fragmentation of the political landscape.

To facilitate the analysis of the case where n goes to infinity but k remains proportional to n ($k \propto n \to \infty$), I define $x_c := n^2\beta a_c$ and $x_{nc} := n^2\beta a_{nc}$, which normalizes abatement levels such that $x_r = 1$ for all regions r in the uncoordinated Nash equilibrium. Expressed in these variables, equations (7)–(9) are

$$x_c = k\left(1 + \frac{1}{x_{nc} - 1 + \mu}\right) \tag{10}$$

$$x_{nc} = 1 + \mu \tag{11}$$

$$\mu = \frac{k(x_c - 1)}{(n-k)(x_{nc} - 1)}. \tag{12}$$

I take the ansatz that in the limit of $k \propto n \to \infty$, x_{nc} behaves like n^d with $d > 0$,[6]. Then, equation (11) implies that μ behaves like n^d as well, and it follows from equation (10) that x_c must behave like n. Equation (12) then implies that μ must behave like n^{1-d}. As μ behaves like both n^d and n^{1-d}, d must be equal to 0.5. Assuming $d \leq 0$ leads to contradictions.

The uncoordinated Nash equilibrium implies that a_r behaves like n^{-2}, and the social optimum would imply that a_r behave like n^{-1} (while the global provision of abatement remains constant, regional contributions shrink with region size). And the outcome of the matching scheme is that the abatement by coalition regions a_c behaves like n^{-1} indeed (coalition regions always internalize the benefits of the whole coalition), but the abatement by non-coalition regions a_{nc} behaves like $n^{-3/2}$. Thus, it is shrinking relative to the social optimum, but at least growing relative to the uncoordinated Nash equilibrium (as the non-

[6] i.e., there exists a real number Ξ, such that $\frac{x_{nc}}{n^d} \to \Xi$.

coalition's contribution shrinks, the coalition can increase the matching rate, counteracting the myopia of the non-coalition).

The behavior thus predicted can indeed be observed for the numerical solution of equations (7)–(9) as given in Figure 2. Note that without any coordination, $x_c = x_{nc} = 1$ for all n (by normalization of x_c and x_{nc}), and under the social optimum, $x_c = x_{nc} = n$. While x_c will behave like n for more general assumptions about abatement cost, the fact that x_{nc} and μ behave like $\sqrt{n}$ depends on the assumption of linearly increasing marginal abatement cost. Appendix B establishes the corresponding results for more general formulations of the abatement cost function.

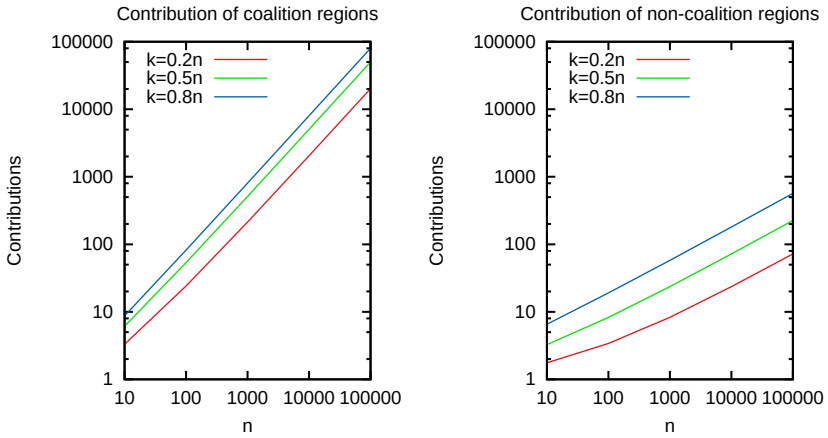

Figure 2: Contribution of coalition and non-coalition regions if coalition covers a constant share of the world while the number of regions varies. Contributions are normalized such that $x_c = x_{nc} = 1$ in the uncoordinated Nash equilibrium.

A.2 Internal and external stability

The literature on coalition formation is concerned about two forms of stability: internal stability, which states that no members want to leave the coalition and external stability, which states that no non-members will join the coalition. As the expansion of coalitions will arguably lead to better global emission outcomes, I regard the expandability of coalitions as a desirable thing and try to analyze which coalitions display this property. I call a coalition expandable, if there is

at least one non-member country that would like to join it.

The welfare of countries be measured in monetary terms. Non-abatement welfare is the benchmark. A county's welfare with respect to the benchmark is the difference between benefits from global abatement and the country's cost of abatement.

$$\begin{aligned} w_{c,(k,n)} &= \frac{A_{(k,n)}}{n} - c_c(a_{c,(k,n)}) \\ &= \frac{ka_{c,(k,n)} + (n-k)a_{nc,(k,n)}}{n} - \frac{n\beta a_{c,(k,n)}^2}{2} \\ w_{nc,(k,n)} &= \frac{A_{(k,n)}}{n} - c_{nc}(a_{nc,(k,n)}) \\ &= \frac{ka_{c,(k,n)} + (n-k)a_{nc,(k,n)}}{n} - \frac{n\beta a_{nc,(k,n)}^2}{2}. \end{aligned}$$

Internal stability requires $w_c(k,n) - w_{nc}(k-1,n) \geq 0$. Asymptotically (as $k \propto n \to \infty$),

$$\begin{aligned} &\frac{ka_{c,(k,n)} + (n-k)a_{nc,(k,n)}}{n} - \frac{n\beta a_{c,(k,n)}^2}{2} \\ &- \frac{(k-1)a_{c,(k-1,n)} + (n-k+1)a_{nc,(k-1,n)}}{n} - \frac{n\beta a_{nc,(k-1,n)}^2}{2} \\ &= \frac{kO(n^{-1}) + (n-k)O(n^{-3/2})}{n} - \frac{n\beta O(n^{-2})}{2} \\ &- \frac{(k-1)O(n^{-1}) + (n-k)O(n^{-3/2})}{n} - \frac{n\beta O(n^{-3})}{2}, \end{aligned}$$

i.e., only the terms involving $a_{c,(k,n)}$ remain relevant as $k \propto n \to \infty$ (if they do not cancel each other out, that is). Looking at only these terms, the condition for internal stability is

$$\begin{aligned} 0 \leq& \frac{ka_{c,(k,n)}}{n} - \frac{n\beta a_{c,(k,n)}^2}{2} - \frac{(k-1)a_{c,(k-1,n)}}{n} \\ \approx& \frac{(k/n)^2/\beta}{n} - \frac{n\beta k^2/(\beta^2 n^4)}{2} - \frac{(k-1)^2/(\beta n^2)}{n} \\ =& \frac{k^2/2 - (k-1)^2}{\beta n^3} \\ =& \frac{2k - (k^2/2+1)}{\beta n^3}, \end{aligned}$$

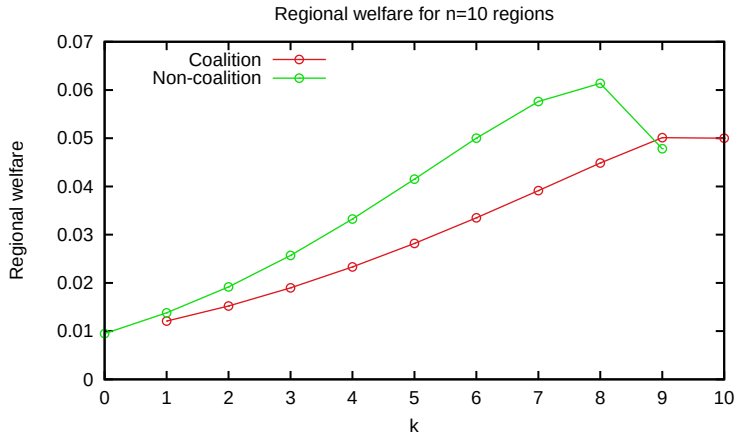

Figure 3: Welfare for countries in- and outside a given matching coalition of size k. Only coalitions of size $k = 1$, $k = 2$, and $k = 10$ are internally stable (their members are better off than non-members in the presence of a coalition of size $k - 1$).

which does not hold in the limit of $k \propto n \to \infty$. As the above is non-zero, restricting the analysis to the terms that vanish no faster than $1/n^2$ is valid. The approximation of $a_{c,(k,n)} \approx k/(\beta n^2)$ is granted by the limiting behavior of equation (7). The results indicate that, as the degree of regional fragmentation increases, no given share of the global population can be kept inside the coalition even if the coalition increases the non-coalition's efforts by matching. The numerical results for welfare w_c and w_{nc} in Figures 3–4 illustrate this trend. The numerical results indicate that for $n = 10, 100, 1000$, only coalitions of size $k = 1, 2, n$ are internally stable ($w_{nc,k-1} < w_{c,k}$). Thus, if coalitions have to be built from scratch ($k = 0$), the share of the global population that can be kept in a stable single coalition shrinks from twenty to two to 0.2 percent as n increases from 10 to 100 to 1000. However, if the threat of reducing emission reduction efforts in proportion to deviations in reductions of defecting countries is used in the context of the full coalition ($k = n$), matching can deter possible free-riders from leaving the coalition.

To reach at the conclusions made so far, I have made two separate simplifications. One simplification is the assumption of symmetry between countries

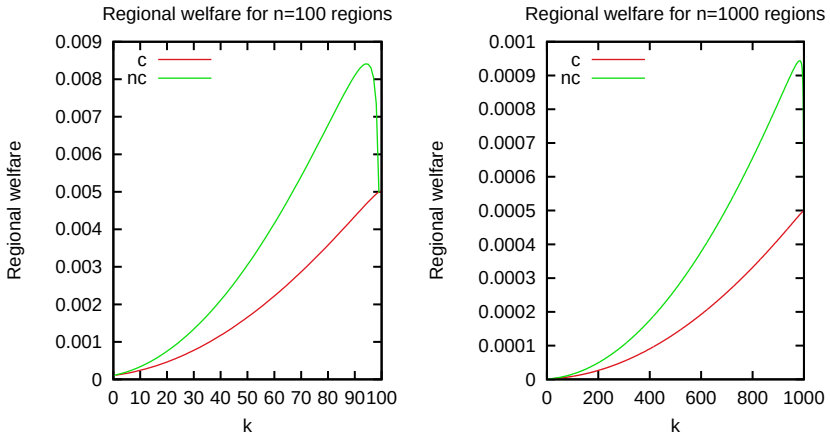

Figure 4: Welfare for countries in- and outside a given matching coalition of size k for $n = 100$ (left panel) and $n = 1000$ (right panel).

playing the public good game, and the other is the quadratic global cost function for providing the public good. It is the latter that drives the results for $k \propto n \to \infty$. In Appendix B, I consider more general cost functions and show that, while increasing the convexity of the cost function decreases the rate at which non-coalition contributions vanish in the limit $k \propto n \to \infty$, no given share of the global population can be kept in the coalition as the political units relevant to the negotiation process become smaller.

B The symmetrical setting with more general abatement cost function

Assume n countries and assume k of them form a coalition. All countries have same constant marginal benefit $mb = 1/n$ from abatement and face an abatement cost curve $c(a) = n^{\gamma}\beta a^{\gamma+1}/(\gamma+1)$, making their marginal abatement cost $mc(a) = \beta(na)^{\gamma}$.[7]

It would be socially optimal for all countries to abate $1/(\beta^{1/\gamma}n)$ units of

[7]Global abatement $A = na$, then causes global abatement cost $C(A) = n \cdot c(A/n) = \beta A^{\gamma+1}/(\gamma+1)$. Thus global abatement cost is independent of n and n can be interpreted as the number of regions that a given world is divided into.

emissions as

$$\sum_{j=1}^{n} \mathrm{mb}_j = 1 = \mathrm{mc}_i \left(\frac{1}{n} \frac{1}{\beta^{1/\gamma}} \right) \qquad \forall i.$$

If all countries act out of self interest, only their own benefits are taken into account and abatement is $1/[n(n\beta)^{1/\gamma}]$:

$$\mathrm{mb}_i = 1/n = \mathrm{mc}_i \left(\frac{1}{n} \frac{1}{(n\beta)^{1/\gamma}} \right) \qquad \forall i.$$

Given the overall matching rate of the coalition μ, the first-order conditions for the coalition c and non-coalition countries nc are

$$\begin{aligned} \beta(na_c)^{\gamma} &= \left(1 + \frac{\alpha}{(n-k)\left(a_{nc} - \frac{1}{n(n\beta)^{1/\gamma}}\right) + \mu\alpha} \right) \frac{k}{n} \\ \beta(na_{nc})^{\gamma} &= \frac{(1+\mu)}{n}. \end{aligned}$$

In agreement with the Section 2, α is determined by

$$\begin{aligned} & n\beta\gamma(na_{nc})^{\gamma-1} n \mathrm{d}a_{nc} = \mathrm{d}(\mu) \\ \Rightarrow \qquad & \alpha = (n-k)\frac{\mathrm{d}a_{nc}}{\mathrm{d}(\mu)} = \frac{(n-k)}{n^2\beta\gamma(na_{nc})^{\gamma-1}}. \end{aligned}$$

Thus, the first-order conditions become

$$n\beta(na_c)^{\gamma} = k\left(1 + \frac{1}{\mu + n\beta\gamma(na_{nc})^{\gamma} - (n\beta)^{1-1/\gamma}(na_{nc})^{\gamma-1}}\right) \tag{13}$$

$$n\beta(na_{nc})^{\gamma} = 1 + \mu. \tag{14}$$

Full use of a_c for matching (threat point is the non-coordination level of abate-

ment) requires

$$k\left(a_c - \frac{1}{n(n\beta)^{1/\gamma}}\right) = \mu(n-k)\left(a_{nc} - \frac{1}{n(n\beta)^{1/\gamma}}\right)$$

$$\Rightarrow \qquad \mu = \frac{k((n\beta)^{1/\gamma} n a_c - 1)}{(n-k)((n\beta)^{1/\gamma} n a_{nc} - 1)} \tag{15}$$

Defining $x_c := n\beta(na_c)^\gamma$ and $x_{nc} = n\beta(na_{nc})^\gamma$,[8] the equations defining the equilibrium are

$$x_c = k\left(1 + \frac{1}{\mu + \gamma\left(x_{nc} - x_{nc}^{1-1/\gamma}\right)}\right)$$

$$x_{nc} = 1 + \mu$$

$$\mu = \frac{k(x_c^{1/\gamma} - 1)}{(n-k)(x_{nc}^{1/\gamma} - 1)}.$$

Using the ansatz that in the limit $n \to \infty$, $x_{nc} \sim n^d$, I conclude that $\mu \sim n^d$ and $x_c \sim n$. Then, $\mu \sim n^{(1-d)/\gamma}$ from the last equation implies $d = (1-d)/\gamma$ and thus $d = \frac{1}{\gamma+1}$.

B.1 Stability

Countries' welfare with respect to the benchmark includes benefits from global abatement as well as local abatement cost.

$$\begin{aligned} w_{c,(k,n)} &= \frac{A_{(k,n)}}{n} - c_c(a_{c,(k,n)}) \\ &= \frac{k a_{c,(k,n)} + (n-k) a_{nc,(k,n)}}{n} - \frac{n^\gamma \beta a_{c,(k,n)}^{\gamma+1}}{\gamma+1} \\ w_{nc,(k,n)} &= \frac{A_{(k,n)}}{n} - c_{nc}(a_{nc,(k,n)}) \\ &= \frac{k a_{c,(k,n)} + (n-k) a_{nc,(k,n)}}{n} - \frac{n^\gamma \beta a_{nc,(k,n)}^{\gamma+1}}{\gamma+1}. \end{aligned}$$

[8] This normalization again implies $x_c = 1, x_{nc} = 1$ as the non-coordination benchmark.

Internal stability requires $w_c(k,n) - w_{nc}(k-1,n) \geq 0$. As $x_c := n\beta(na_c)^\gamma$ goes like n,

$$a_c = O(n^{-1}),$$

and as x_{nc} goes like $n^{1/(1+\gamma)}$,

$$a_{nc} = O(n^{[1/(1+\gamma)-(1-\gamma)]/\gamma}) = O(n^{-[1+1/(\gamma^2+\gamma)]}).$$

Asymptotically (as $k \propto n \to \infty$),

$$\begin{aligned}
&\frac{ka_{c,(k,n)} + (n-k)a_{nc,(k,n)}}{n} - \frac{n^\gamma \beta a_{c,(k,n)}^{\gamma+1}}{\gamma+1} \\
&- \frac{(k-1)a_{c,(k-1,n)} + (n-k+1)a_{nc,(k-1,n)}}{n} - \frac{n^\gamma \beta a_{nc,(k-1,n)}^{\gamma+1}}{\gamma+1} \\
&= \frac{kO(n^{-1}) + (n-k)O(n^{-[1+1/(\gamma^2+\gamma)]})}{n} - \frac{n^\gamma \beta O(n^{-(1+\gamma)})}{\gamma+1} \\
&- \frac{(k-1)O(n^{-1}) + (n-k)O(n^{-[1+1/(\gamma^2+\gamma)]})}{n} - \frac{n^\gamma \beta O(n^{-[(1+\gamma)+(1+\gamma)/(\gamma^2+\gamma)]})}{\gamma+1},
\end{aligned}$$

i.e., only the terms involving $a_{c,(k,n)}$ remain relevant as $k \propto n \to \infty$ (if they do not cancel each other out, that is). Looking at only these terms, the condition for internal stability is

$$\begin{aligned}
0 \leq& \frac{ka_{c,(k,n)}}{n} - \frac{n^\gamma \beta a_{c,(k,n)}^{1+\gamma}}{1+\gamma} - \frac{(k-1)a_{c,(k-1,n)}}{n} \\
\approx& \frac{k}{n}\frac{1}{n}\left(\frac{k}{\beta n}\right)^{1/\gamma} - \frac{n^\gamma \beta \left(\frac{1}{n}\left[\frac{k}{\beta n}\right]^{1/\gamma}\right)^{1+\gamma}}{1+\gamma} - \frac{(k-1)}{n}\frac{1}{n}\left(\frac{k-1}{\beta n}\right)^{1/\gamma} \\
=& \frac{1}{n\beta^{1/\gamma}}\left(\frac{k}{n}\right)^{1+1/\gamma} - \frac{\frac{1}{n\beta^{1/\gamma}}\left(\frac{k}{n}\right)^{1+1/\gamma}}{1+\gamma} - \frac{1}{n\beta^{1/\gamma}}\left(\frac{k-1}{n}\right)^{1+1/\gamma} \\
=& \frac{1}{n\beta^{1/\gamma}}\left[\left(\frac{k}{n}\right)^{1+1/\gamma}\left(1 - \frac{1}{1+\gamma}\right) - \left(\frac{k-1}{n}\right)^{1+1/\gamma}\right] \\
\overset{k\propto n\to\infty}{\to}& \frac{1}{n\beta^{1/\gamma}}\left[\left(\frac{k}{n}\right)^{1+1/\gamma}\left(-\frac{1}{1+\gamma}\right)\right].
\end{aligned}$$

The approximation of $a_{c,(k,n)} \approx 1/n \cdot [k/(\beta n)]^{1/\gamma}$ is granted by the limiting behavior of equation (13).

As this the right hand side of this condition is clearly lower than 0, we see that no share of total population can be held inside the coalition as the political units at the negotiating table shrink.

C Uniqueness

Assume the asymmetrical case $\mathrm{mc}_r(a_r) = c_r a_r^\gamma$, $\mathrm{mb}_r = b_r$. In an uncoordinated Nash equilibrium, all countries would abate amounts $\hat{a}_r = (b_r/c_r)^{1/\gamma}$. The first-order conditions of the coalition and non-coalition countries are

$$\begin{aligned} c_{nc} a_{nc}^\gamma &= b_r \, (1+\mu) \\ c_c a_c^\gamma &= b_r \left(k + \frac{k}{\mu + \frac{\sum_{nc}(a_{nc} - \hat{a}_{nc})}{\alpha}} \right) \end{aligned}$$

and can be solved to

$$a_{nc} = \hat{a}_{nc}(1+\mu)^{1/\gamma} \tag{16}$$

$$\begin{aligned} a_c &= \hat{a}_c \left(k + \frac{k}{\mu + \frac{\sum_{nc}(a_{nc} - \hat{a}_{nc})}{\alpha}} \right)^{1/\gamma} \\ &= \hat{a}_c \left(k + \frac{k}{\mu + \underbrace{\frac{((1+\mu)^{1/\gamma} - 1) \sum_{nc} \hat{a}_{nc}}{\frac{(1+\mu)^{1/\gamma - 1}}{\gamma} \sum_{nc} \hat{a}_{nc}}}_{\alpha}} \right)^{1/\gamma} \\ &= \hat{a}_c \left(k + \frac{k}{\mu + \gamma(1+\mu)(1 - [1+\mu]^{-1/\gamma})} \right)^{1/\gamma}. \end{aligned} \tag{17}$$

Equations (16) and (17) imply that a_{nc} are strictly increasing and a_c are strictly decreasing in $\mu \geq 0$. But μ is set according to

$$\sum_c (a_c - \hat{a}_c) = \mu \sum_{nc} (a_{nc} - \hat{a}_{nc}). \tag{18}$$

If the solutions to (16) and (17) are inserted, the right hand side of equation (18) become strictly increasing while its left hand side become strictly decreasing. For $\mu \to 0$, the left hand side is greater than the right hand side, and for $\mu \to \infty$, the left hand side is smaller than the right hand side. All this together implies that there must be a unique μ and corresponding a_cs and a_{nc}s solving the equations (16)–(18).