UvA

Multi-Agent Systems

Coalition Logic

Mina Young Pedersen
m.y.pedersen@uva.nl

Fall 2025

What is Coalition Logic?

- We have seen the basics of modal logic
    - But what can we use it for?
- **Coalition Logic (CL)** is a type of **strategy logic**
    - A **strategy** is a conditional plan
        - Intuitively, a strategy is intended to work whatever the opponents (other agents, environment) do
        - Coalition Logic looks at *one-step* strategies:
            - Is there an action by a coalition of agents *C* such that whatever action other agents perform, *C* “win” (make sure some property holds)?

Why Coalition Logic?

- A formal system to analyze the strategies and abilities of a group/coalition of agents
- It allows us to reason about voting situations
    - Therefore, it is a great example of a system at the intersection of modal logic and voting theory

Coalition Models

- We begin by introducing our semantic structures
- As was the case in the basic modal logic, we define a set At of atoms

Coalition Models

Definition (Coalition Model)

Let N be a set of agents. A **coalition model** is a tuple $M = (W, E, V)$, where:

- W is a non-empty domain (whose elements are generically called **(possible) worlds** or **states**);
- $E : \mathscr{P}(N) \to \mathscr{P}(\mathscr{P}(W))$ is the **effectivity function** (assigning, to each set of agents in $\mathscr{P}(N)$, a set of sets of worlds);
- $V : \mathrm{At} \to \mathscr{P}(W)$ is an atomic **valuation** (indicating the set of worlds that satisfy each one of the atomic propositions).

For M to be a coalition model, the effectivity function E also needs to satisfy the following conditions. For *every coalition* $C \subseteq N$, we should have both:

$$\emptyset \notin E(C) \text{ and } W \in E(C)$$

More on Coalition Models

- A coalition model contains both a (non-empty) set of worlds W and a valuation V, just as a relational model
    - The difference between them is that a coalition model contains an effectivity function E instead of a relation R

- The models are intended to represent the possible outcomes of some interaction among the agents
    - And the effectivity function specifies the abilities each coalition has
    - More precisely, $U \in E(C)$ is read as "*coalition C can ensure that the final outcome will be among the worlds in U*"

Example

$$W = \{w_1, w_2, w_3, w_4\}$$

$$V(p) = \{w_1, w_2\}$$

$$V(q) = \{w_1, w_3\}$$

$$N = \{1,2\}$$

We must assign a set of sets of worlds to each coalition $C \in \mathscr{P}(N)$

$$E(\varnothing) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{2\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1,2\}) = \{\{w_1, w_2\}, \{w_1, w_2, w_3, w_4\}\}$$

Example

$$E(\emptyset) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{2\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1,2\}) = \{\{w_1, w_2\}, \{w_1, w_2, w_3, w_4\}\}$$

What can we say about this model?

- The effectivity function defines a proper coalition model: For every coalition $C \subseteq N$, we have $\emptyset \notin E(C)$ and $W \in E(C)$
- The coalitions $\emptyset$, $\{1\}$ and $\{2\}$ essentially have no power: they can only guarantee that the outcome will be among the worlds in W

Example

$$E(\emptyset) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{2\}) = \{\{w_1, w_2, w_3, w_4\}\}$$
$$E(\{1,2\}) = \{\{w_1, w_2\}, \{w_1, w_2, w_3, w_4\}\}$$

What can we say about this model?

- The coalition $\{1,2\}$ can guarantee that the outcome will be in $\{w_1, w_2\}$
  - Although the agents 1 and 2 have no power on their own
- Since p is true in all worlds in $\{w_1, w_2\}$, we can actually say that coalition $\{1,2\}$ can guarantee that p will be the case

Language of Coalition Logic

- In the previous example, we saw that the coalition $\{1,2\}$ has the power/ability to guarantee that p will be the case
- Adding a formal language will help us express these and similar ideas

Definition (Coalition Language)

Let N be a set of agents, and At be a set of atomic propositions. Formulas ϕ, ψ of the **coalition language** are given inductively as follows:

$$\phi, \psi ::= \text{At} \mid \neg\phi \mid (\phi \wedge \psi) \mid \langle\langle C \rangle\rangle \phi$$

with $C \subseteq N$. We define $[[C]]\phi := \neg\langle\langle C \rangle\rangle \neg\phi$ and ($\vee$, $\rightarrow$, $\leftrightarrow$, $\top$, $\bot$) as standard.

Notes on the Coalition Language

- Formulas on the form $\langle\langle C \rangle\rangle \phi$ are read as "*the coalition C can cooperate to ensure that* ϕ *holds*"
- As before, we define:
  - $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$
  - $\phi \rightarrow \psi := \neg(\phi \wedge \neg\psi)$
  - $\phi \leftrightarrow \psi := (\phi \rightarrow \psi) \wedge (\psi \rightarrow \phi)$
  - $\top := p \vee \neg p$
  - $\bot := p \wedge \neg p$

Semantics

- With the semantics, we connect the coalition models and the coalition language

Definition (Semantic Interpretation/Truth)

Let $M = (W, E, V)$ be a coalition model and $w \in W$ be a world.

Truth of a coalition formula ϕ at world w in M, written $M, w \Vdash \phi$, is defined inductively as follows:

$M, w \Vdash p$ **iff** $w \in V(p)$

$M, w \Vdash \neg\phi$ **iff** $M, w \nVdash \phi$

$M, w \Vdash \phi \wedge \psi$ **iff** $M, w \Vdash \phi$ **and** $M, w \Vdash \psi$

$M, w \Vdash \langle\langle C \rangle\rangle \phi$ **iff there is** $U \in E(C)$ **such that**

for all $u \in W : u \in U$ **implies** $M, u \Vdash \phi$

Notes on the Semantics

- Formulas are evaluated in *pointed* models
    - In some sense, the language takes a *local* point of view
    - But our modal operator $\langle\langle C \rangle\rangle \phi$ works *globally*: if it is true, it is true at all points in the model
- The $\langle\langle C \rangle\rangle \phi$ operator works as a *sequence* of an existential quantification followed by a universal one
    - It requires the *existence* of a set of worlds that the coalition can “enforce” with *all* the worlds in this set satisfying the given formula
    - In other words, the coalition *C* can cooperate to ensure ϕ (i.e., $\langle\langle C \rangle\rangle \phi$) if and only if *C* can guarantee that the outcome will be in a set that contains only worlds in which ϕ is the case

Notes on the Semantics

- Recall that we defined $[[C]]\phi := \neg\langle\langle C\rangle\rangle\neg\phi$
- As you can verify:

  $M, w \Vdash [[C]]\phi$ for all $U \in E(C)$ there is $u \in W$ such that $u \in U$ and $M, u \Vdash \phi$

  - Thus, the $[[C]]\phi$ is a sequence of a *universal* quantification and then an *existential* one

- $[[C]]\phi$ can be read as "*coalition C cannot ensure that ϕ fails*"

Example

- Taken from Gibbard (1974)*:

> If *Angelina* (*a*) does not want to remain single, she can decide to marry *Edwin* (*e*) or *the judge* (*j*). Edwin and the judge each can similarly decide whether they want to stay single or marry Angelina. Assume the three individuals live in a society where nobody can be forced to marry against her/his will.

*A. Gibbard (1974): A Pareto-Consistent Libertarian Claim. *Journal of Economic Theory*, 7(4):388—410.

Example

- This is a situation where agents can cooperate to ensure different outcomes
- Let’s use Coalition Logic to reason about Angelina’s marital status
- Given that she can be either single, or else married to Edwin, or else married to the judge, it makes sense to use three atomic propositions p_s, p_e, p_j
- Also, it makes sense to use three possible worlds, each one representing a situation where one proposition is true and the other two are false

Example

$$W = \{w_s, w_e, w_j\}$$

$$N = \{a, e, j\}$$

$$V(p_s) = w_s$$

$$V(p_e) = w_e$$

$$V(p_j) = w_j$$

We can think of the members of $\mathscr{P}(W)$ as sets of outcomes.

But what does it mean for the outcome to be in each set of possible worlds?

$\{w_s\}$: "*a* is single"

$\{w_e\}$: "*a* and *e* are married to each other"

$\{w_j\}$: "*a* and *j* are married to each other"

Example

$$W = \{w_s, w_e, w_j\}$$

$$N = \{a, e, j\}$$

$$V(p_s) = w_s$$

$$V(p_e) = w_e$$

$$V(p_j) = w_j$$

$\{w_s, w_e\}$: “*a* and *j* are *not* married to each other”

$\{w_s, w_j\}$: “*a* and *e* are *not* married to each other”

$\{w_e, w_j\}$: “*a* is married”

$\{w_s, w_e, w_j\}$: “*a* is either married or else single”

Example

$$W = \{w_s, w_e, w_j\}$$

$$N = \{a, e, j\}$$

$$V(p_s) = w_s$$

$$V(p_e) = w_e$$

$$V(p_j) = w_j$$

We can now define the effectivity function.

$$E(\{a\}) := \{\{w_s\}, \{w_s, w_e\}, \{w_s, w_j\}, \{w_s, w_e, w_j\}\}$$

$$E(\{e\}) := \{\{w_s, w_j\}, \{w_s, w_e, w_j\}\}$$

$$E(\{j\}) := \{\{w_s, w_e\}, \{w_s, w_e, w_j\}\}$$

$$E(\{a, e\}) := \{\{w_s\}, \{w_e\}, \{w_s, w_e\}, \{w_s, w_j\}, \{w_e, w_j\}, \{w_s, w_e, w_j\}\}$$

$$E(\{a, j\}) := \{\{w_s\}, \{w_j\}, \{w_s, w_e\}, \{w_s, w_j\}, \{w_e, w_j\}, \{w_s, w_e, w_j\}\}$$

$$E(\{e, j\}) := \{\{w_s\}, \{w_s, w_e\}, \{w_s, w_j\}, \{w_s, w_e, w_j\}\}$$

$$E(\{a, e, j\}) := \mathscr{P}(W) \backslash \varnothing$$

We left out the case for $C = \varnothing$

Example

$$W = \{w_s, w_e, w_j\}$$

$$N = \{a, e, j\}$$

$$V(p_s) = w_s$$

$$V(p_e) = w_e$$

$$V(p_j) = w_j$$

Here are some things we can express with coalition formulas for all $w \in W$:

$$M, w \Vdash \langle\langle\{a\}\rangle\rangle p_s \wedge \langle\langle\{a\}\rangle\rangle(p_s \vee p_j) \wedge \langle\langle\{a\}\rangle\rangle(p_s \vee p_e)$$

a has the power to remain single, the power to *not* marry *e*, and the power *not* to marry *j*

$$M, w \Vdash \langle\langle\{e\}\rangle\rangle(p_s \vee p_j) \wedge [[\{e\}]]p_j$$

e has the power to guarantee that *a* does not marry him, but cannot ensure that *a* and *j* will not be married

$$M, w \Vdash \neg\langle\langle\{a\}\rangle\rangle p_e \wedge \neg\langle\langle\{e\}\rangle\rangle p_e \wedge \langle\langle\{a, e\}\rangle\rangle p_e$$

neither *a* nor *e* can guarantee on their own that they will marry each other, but if they work together, they can guarantee it

Example: Finding a Model

- Taken from Pauly (2001)*
- Consider the following scenario:

> Two individuals, 1 and 2, are to choose between two options, *a* and *b*. We want a procedure for making the choice that will satisfy the following requirements:
>
> 1. First, we want for both options to be possible—that is, it should be possible for the coalition of both agents to bring about *a*, and it should also be possible for their coalition to bring about *b*.
>
> 2. We do not want them to be able to bring about both options simultaneously.
>
> 3. Similarly, we do not want either agent to dominate: we want them both to have equal power.

*M. Pauly (2001): *Logic for Social Software*. PhD thesis, ILLC, UvA

Example: Finding a Model

- Is it possible to satisfy these requirements? If so, how?
    - Let's write down the requirements as logical formulas

1. Both options should be possible

$$\langle\langle\{1,2\}\rangle\rangle a \wedge \langle\langle\{1,2\}\rangle\rangle b$$

2. They should not be able to bring about both options simultaneously

$$\neg\langle\langle\{1,2\}\rangle\rangle(a \wedge b)$$

3. They should both have equal power

$$(\neg\langle\langle\{1\}\rangle\rangle a \wedge \neg\langle\langle\{1\}\rangle\rangle b) \wedge (\neg\langle\langle\{2\}\rangle\rangle a \wedge \neg\langle\langle\{2\}\rangle\rangle b)$$

Now, the problem of finding how to satisfy the requirements becomes finding a coalition model where all these formulas are true…

Example: Finding a Model

$\langle\langle\{1,2\}\rangle\rangle a \wedge \langle\langle\{1,2\}\rangle\rangle b$

$\neg\langle\langle\{1,2\}\rangle\rangle(a \wedge b)$

$(\neg\langle\langle\{1\}\rangle\rangle a \wedge \neg\langle\langle\{1\}\rangle\rangle b) \wedge (\neg\langle\langle\{2\}\rangle\rangle a \wedge \neg\langle\langle\{2\}\rangle\rangle b)$

$N = \{1,2\}$

$W = \{w_1, w_2, w_3, w_4\}$

All possible combinations of truth values for *a* and *b*

$\{a, b\}$

$V(a) = \{w_1, w_2\}$

$V(b) = \{w_1, w_3\}$

Next: Define the effectivity function!

Example: Finding a Model

$$\langle\langle\{1,2\}\rangle\rangle a \wedge \langle\langle\{1,2\}\rangle\rangle b$$

$$\neg\langle\langle\{1,2\}\rangle\rangle(a \wedge b)$$

$$(\neg\langle\langle\{1\}\rangle\rangle a \wedge \neg\langle\langle\{1\}\rangle\rangle b) \wedge (\neg\langle\langle\{2\}\rangle\rangle a \wedge \neg\langle\langle\{2\}\rangle\rangle b)$$

For coalition $\{1,2\}$:

We want $\langle\langle\{1,2\}\rangle\rangle a$:

$E(\{1,2\})$ need at least one set containing only *a*-worlds:
$\{w_1\}, \{w_2\}$ or $\{w_1, w_2\}$

We want $\langle\langle\{1,2\}\rangle\rangle b$:

$E(\{1,2\})$ need at least one set containing only *a*-worlds:
$\{w_1\}, \{w_3\}$ or $\{w_1, w_3\}$

We want $\neg\langle\langle\{1,2\}\rangle\rangle(a \wedge b)$:

No set in $E(\{1,2\})$ contains only $(a \wedge b)$-worlds: $\{w_1\}$ should not be in

+ we need to satisfy the conditions $\emptyset \notin E(\{1,2\})$ and $W \in E(\{1,2\})$

One alternative: $E(\{1,2\}) = \{\{w_2\}, \{w_3\}, W\}$

Example: Finding a Model

$$\langle\langle\{1,2\}\rangle\rangle a \wedge \langle\langle\{1,2\}\rangle\rangle b$$

$$\neg\langle\langle\{1,2\}\rangle\rangle(a \wedge b)$$

$$(\neg\langle\langle\{1\}\rangle\rangle a \wedge \neg\langle\langle\{1\}\rangle\rangle b) \wedge (\neg\langle\langle\{2\}\rangle\rangle a \wedge \neg\langle\langle\{2\}\rangle\rangle b)$$

For coalition $\{1\}$:

We want $\neg\langle\langle\{1\}\rangle\rangle a$:

We need that no set in $E(\{1\})$ contain only *a*-worlds: neither $\{w_1\}, \{w_2\}$ nor $\{w_1, w_2\}$ should be in

We want $\neg\langle\langle\{1\}\rangle\rangle b$:

We need that no set in $E(\{1\})$ contain only *b*-worlds: neither $\{w_1\}, \{w_3\}$ nor $\{w_1, w_3\}$ should be in

+ we need to satisfy the conditions $\emptyset \notin E(\{1,2\})$ and $W \in E(\{1,2\})$

One alternative: $E(\{1\}) = \{W\}$

Similar reasoning for the coalition $\{2\}$ yields: $E(\{2\}) = \{W\}$

We also define $E(\emptyset) = \{W\}$

Finding a Class of Models

- In some cases, one is interested in finding a coalition model satisfying certain specific requirements

    - In some others, one might rather be interested in finding a *class* of coalition models in which the “abilities” of all coalitions have some *general* properties

Example: Finding a Class of Models

- Let N be a set of agents and $C \subseteq N$ be an arbitrary coalition
- Until now, we have looked at concrete formulas indicating particular “powers” in a specific model
- Now, we look at general *formula schema* indicating general properties the abilities of all coalitions should have in the *collection* of models we want to work with

Example: Finding a Class of Models

- For example, we might be interested in:

  Coalition models where every coalition has the power to ensure contradictions

First thought: $\langle\langle C \rangle\rangle \bot$ should be always true

By the semantics $\langle\langle C \rangle\rangle \bot$ is true when in $E(C)$ there is a set that contains only worlds where $\bot$ is true

But $\bot$ is false everywhere

So we need $\emptyset \in E(C)$

But this is forbidden by the definition of the coalition model

It is not possible to make $\langle\langle C \rangle\rangle \bot$ true

Example: Finding a Class of Models

- What if we are instead interested in:

  Coalition models where, if a coalition has the ability to enforce an implication, and also its antecedent, then it also has the ability to enforce the consequent

What formula (schema) expresses this property?

$$(\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle C\rangle\rangle\phi) \to \langle\langle C\rangle\rangle\psi$$

Find a class of coalition models where this formula is always true:

We only need to look at models where the antecedent $\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle C\rangle\rangle\phi$ is true

Example: Finding a Class of Models

$$(\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle \{C\}\rangle\rangle\phi) \to \langle\langle \{C\}\rangle\rangle\psi$$

Find a class of coalition models where this formula is always true:

Take an arbitrary model where the antecedent $\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle C\rangle\rangle\phi$ is true

In this model, fix an arbitrary C

$E(C)$ contains two sets of worlds V_1 and V_2 such that

V_1 contains only worlds where $\phi \to \psi$ is true

V_2 contains only worlds where ϕ is true

We want to have $V_3 \in E(C)$ that contains only worlds where ψ is true

The intersection $V_1 \cap V_2$ contains only worlds where ψ is true

If $E(C)$ is closed under intersections, we have what we want

Example: Finding a Class of Models

$$(\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle\{C\}\rangle\rangle\phi) \to \langle\langle\{C\}\rangle\rangle\psi$$

Find a class of coalition models where this formula is always true:

$(\langle\langle C\rangle\rangle(\phi \to \psi) \wedge \langle\langle\{C\}\rangle\rangle\phi) \to \langle\langle\{C\}\rangle\rangle\psi$ is true (for any coalition C and any formulas ϕ, ψ) in all coalition models where the effectivity function (for every coalition C) is closed under intersections

Additional Conditions on Coalition Models

- We have that in a proper coalition model, every coalition can guarantee that the outcome will be in the domain: $W \in E(C)$ for all $C \subseteq N$
    - And that no coalition can guarantee that there will be no outcome: $\emptyset \notin E(C)$
- But there are also other possible conditions we can put on coalition models

Monotonicity and Super-Additivity

Definition (Monotonicity and Super-Additivity)

Let $M = (W, E, V)$ be a coalition model. It is said that M satisfies:

- **Monotonicity** if and only if, for all $C \subseteq N$, if $U \in E(C)$ and $U \subseteq U'$, then $U' \in E(C)$

Intuitively, M satisfies monotonicity if and only if coalitions can always enforce weaker outcomes.

- **Super-additivity** if and only if for all $C_1, C_2 \subseteq N$ with $C_1 \cap C_2 = \emptyset$, if $U_1 \in E(C_1)$ and $U_2 \in E(C_2)$, then $U_1 \cap U_2 \in E(C_1 \cup C_2)$

Intuitively, M satisfies super-additivity if and only if coalitions can combine their abilities to (possibly) achieve more.

Additional Reading Material

- For additional reading material, I advise:
    - M. Pauly (2001): *Logic for Social Software*. PhD thesis, ILLC, UvA