Argumentation-based Explanation

Xiuyi Fan[1], Francesca Toni[2]

1. Computer Science Department, Swansea University
2. Department of Computing, Imperial College London

Explanation Semantics

Main References

- Xiuyi Fan, Francesca Toni: On Computing Explanations in Argumentation. AAAI 2015: 1496-1502
- Xiuyi Fan, Francesca Toni: On Explanations for Non-Acceptable Arguments. TAFA 2015: 112-127

Argumentation Semantics & Explanations

Argumentation semantics give criteria for "acceptable" arguments.

- Good for answering: *Given a set of arguments, which subsets are "good"?*
- Not good for: *Given a set of arguments, why is a particular argument "good"?*

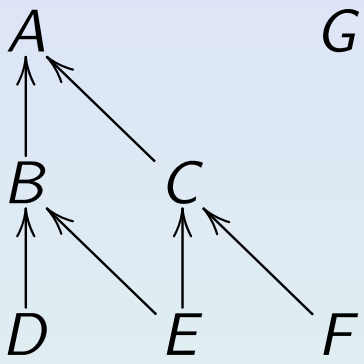
It is widely acknowledged that an explanation should be a *justification* (Newton-Smith81):

> . . . if I am asked to explain why I hold some general belief that p, I answer by giving my justification for the claim that p is true.

Hence, if a belief q does not contribute to the justification of p, q should not be in the explanation of p.

Explanation Example

Hotel selection problem:



A: Choose ic.
B: Why not jh?
D: Because it is not quiet.
C: Why not ritz?
E: But ritz is not cheap (neither is jh).
F: Also, ritz is fully booked.
G: London has good public transport.

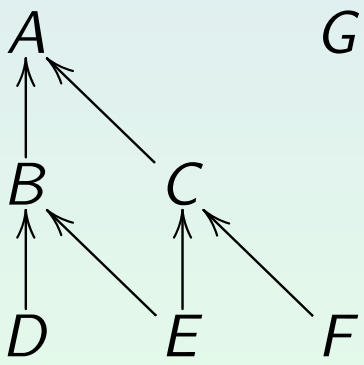
- D, E and F defend A. Hence, together, they fully *justify* A. E by itself or D and F together also *justify* A.
- G has nothing to do with A. Hence, if one is interested in explaining A, G should not be included.

Related Admissibility (I)

Definition

Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, let $X, Y \in \mathcal{A}$. X *defends* Y iff:
1. $X = Y$; or
2. $\exists Z \in \mathcal{A}$, s.t. X attacks Z and Z attacks Y; or
3. $\exists Z \in \mathcal{A}$, s.t. X defends Z and Z defends Y.

$S \subseteq \mathcal{A}$ *defends* $X \in \mathcal{A}$ iff $\forall Y \in S$: Y defends X.



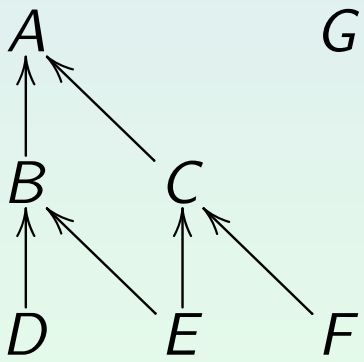
- Every argument defends itself.
- Each of A, D, E and F defends A, and $\{A, D, E, F\}$ and all its non-empty subsets defend A.
- No argument defends G except G itself.

Related Admissibility (II)

"Defends" + "Admissibility" = Related Admissibility

Definition

Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, $S \subseteq \mathcal{A}$ is *related admissible* iff $\exists X \in S$ s.t. S defends X and S is admissible. Any such X is referred to as a *topic* of S.



- $\{A, D, E, F\}$, $\{A, D, E\}$, $\{A, D, F\}$, $\{A, E, F\}$, and $\{A, E\}$ are related admissible, with A the topic of all.
- $\{F, G\}$ is admissible but not related admissible, since F and G do not defend one another.

Explanations

Definition

Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, for any argument $X \in \mathcal{A}$, an *explanation* of X is $S \subseteq \mathcal{A}$ s.t. S is a related admissible set and X is a topic of S.

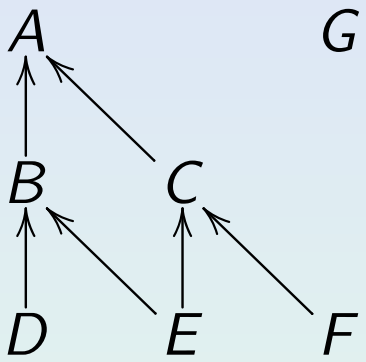
We can classify explanations into different types:

Definition

Given an AA framework $\langle \mathcal{A}, \mathcal{R} \rangle$, let $A \in \mathcal{A}$ and $E_A = \{S | S$ is an explanation of $A\}$. Then, for any $S \in E_A$, S is

- a *Minimal Explanation (MiE)* iff S is smallest wrt. $<$;
- a *Compact Explanation (CE)* iff S is smallest wrt. $\subset$;
- a *Maximal Explanation (MaE)* iff S is largest wrt. $<$;
- a *Verbose Explanation (VE)* iff S is largest wrt. $\subset$.

Explanation



- $\{A, D, E, F\}$ is both a MaE and a VE.
- Both $\{A, D, F\}$ and $\{A, E\}$ are CEs.
- $\{A, E\}$ is a MiE.

Their natural language reading is:

- $\{A, E\}$: choose ic as both jh and ritz are not cheap.
- $\{A, D, F\}$: choose ic as jh is not quiet and ritz is booked.
- $\{A, D, E, F\}$: choose ic for all reasons above.

Dispute Trees

Dispute tree $\mathcal{T}$ for an argument a (informally):

- every node of $\mathcal{T}$ is either $[\mathtt{P}:x]$ or $[\mathtt{O}:x]$ (x is an argument);
- the root of $\mathcal{T}$ is $[\mathtt{P}:a]$;
- children attack their parents;
- a $\mathtt{P}$ node has as many children as arguments attacking it;
- an $\mathtt{O}$ node has at most one child.

A dispute tree $\mathcal{T}$ is admissible iff

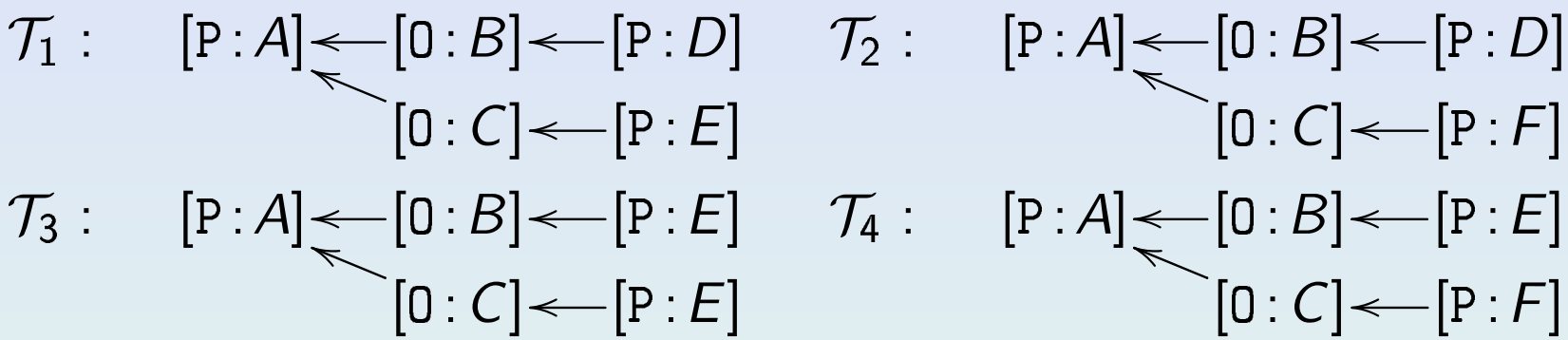
- every $\mathtt{O}$ node in $\mathcal{T}$ has a child;
- no argument in $\mathcal{T}$ labels both a $\mathtt{P}$ and an $\mathtt{O}$ node.

Admissible arguments correspond to admissible dispute trees:

- the argument in the root of an admissible tree is admissible.

Computing Explanations

MiE, CE, MaE and VE are all computed with *dispute forests* - multiple dispute trees.



Results:

- Arguments in P nodes from a tree are related admissible.
- "Smaller" trees give "smaller" explanations.
- "Compatible" tree set gives "larger" explanations.

Related Admissibility in ABA

- Assumption-based Argumentation (ABA) is a structured argumentation framework with *rules, assumptions* and *contraries* (of assumptions) defined over a *language* made of *sentances*.
- In ABA, admissibility is defined both in the level of arguments and assumptions. Thus, by defining the *defend* relation between arguments and sentances, *related admissibility* can be defined in the level of assumptions as well.
- Dispute trees / forests can still be used to compute related admissible sets in ABA.

Conclusion

- Explanation
- Related Admissibility in AA
- Dispute Forests
- Related Admissibility in ABA

Questions?

Key Aspects - Explain Non-acceptable Arguments

Question:

> *Given an AA framework F and some non-admissible argument a in F, why isn't a admissible?*

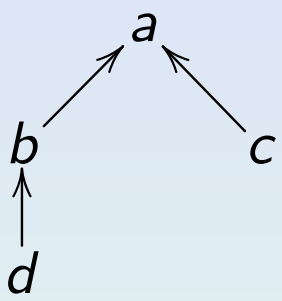
(Here, an argument is admissible if it is in an admissible extension.)

Two different notions of explanation and their computation

- Notions of explanation:
  - Explanation in argument view: *arg-explanation*
  - Explanation in attack view: *att-explanation*
- Explanation computation: prune dispute trees

Example

An agent is choosing a hotel, there are three candidates: ic, jh, ritz.



a: Choose ic.
b: Why not jh?
d: Because it is not quiet.
c: Why not ritz?

Argument view: a is not admissible, because of c.
Attack view: a is not admissible, because of (c, a).

Note that although b also attacks a, but it is "taken care of" by d.

Two notions of Explanation

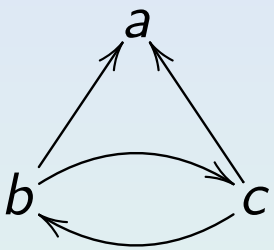
Arg-explanation:

> *For a (non-admissible)* topic *argument a in some AA framework F, an arg-explanation for a is a minimum set of arguments S s.t. if S is removed from F, then a becomes admissible.*

Att-explanation:

> *A minimum set of attacks E s.t. if E is removed from F, then a becomes admissible.*

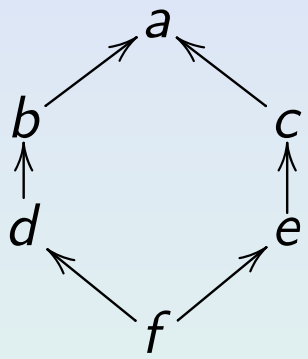
Example 2: arg-explanation vs. att-explanation



a is not admissible as it is attacked by both b and c. To make a admissible, we can

- either remove **both** b and c (as removing only one of them is insufficient) hence the arg-explanation for a is $\{b, c\}$;
- or we can remove **either** the attack (b, a) **or** the attack (c, a).

Example 3: arg-explanation vs. att-explanation



Here, a is not admissible.

- $\{(b, a), (f, e)\}$ is an att-explanation for a.
- It is easy to see that $\{f\}$ is an arg-explanation.

Compute Explanations

- Admissible arguments correspond to admissible dispute trees:
  - the argument in the root of an admissible tree is admissible.
- Non-admissible arguments correspond to non-admissible dispute trees.
- Non-admissible trees can be "turned into" admissible trees.

Question:

How do we turn a non-admissible dispute tree into an admissible dispute tree by removing its nodes?

Compute Arg-Explanation - Example

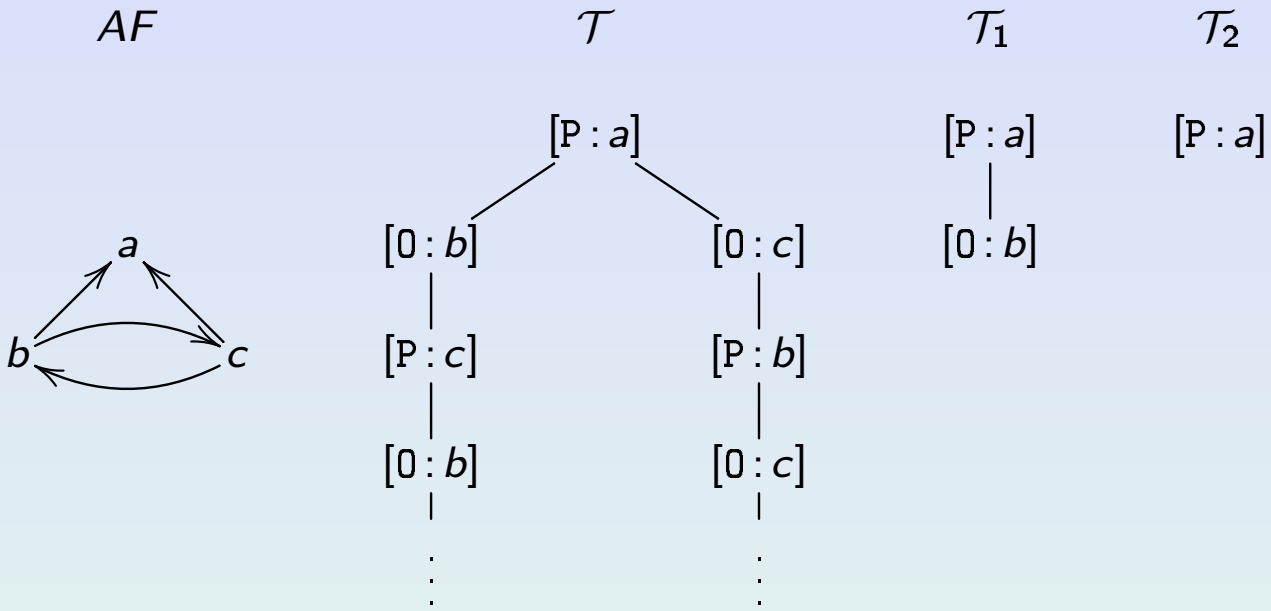


Figure: $\mathcal{T}_1 = \mathcal{T} \setminus \{c\}$; $\mathcal{T}_2 = \mathcal{T} \setminus \{c, b\}$.

For a set of arguments A, $\mathcal{T} \setminus A = \mathcal{T}'$: $\mathcal{T}'$ does not contain any node in A that "hung below" arguments in A.

- $\mathcal{T}$ is not admissible; $\mathcal{T}_1$ is not admissible; $\mathcal{T}_2$ is admissible.
- $\{c, b\}$ is an arg-explanation for a.

Compute Att-Explanation - Example

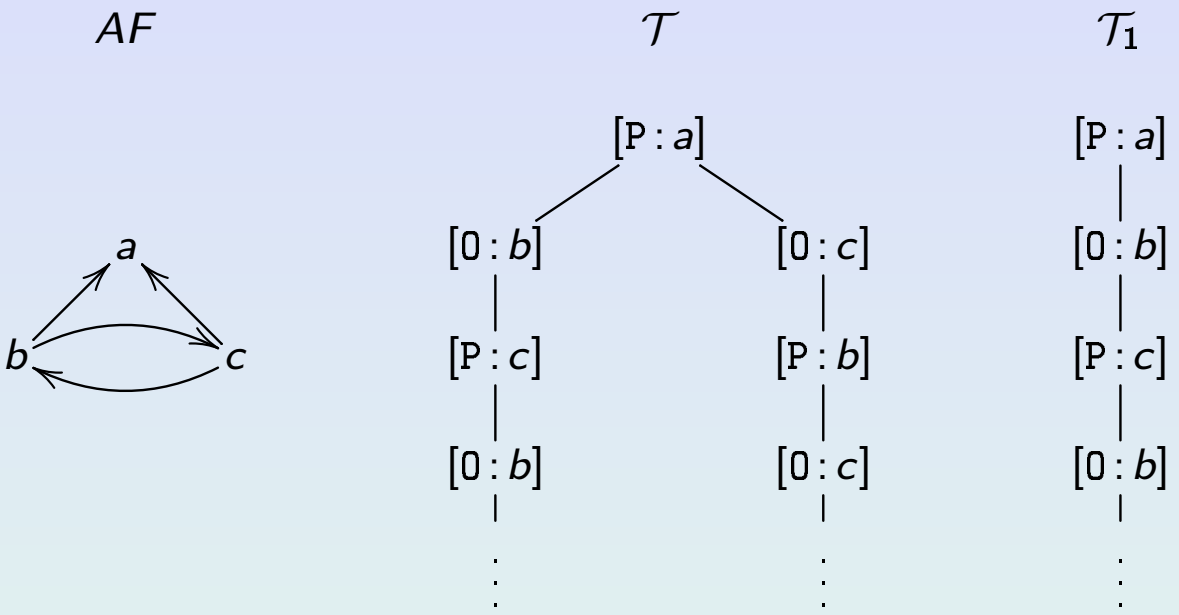


Figure: $\mathcal{T}_1 = \mathcal{T} \setminus \{(c, a)\}$

For a set of attacks E, $\mathcal{T} \setminus E = \mathcal{T}'$: $\mathcal{T}'$ does not contain any branch rooted at x with y the parent of x, where $(x, y) \in E$.

- $\mathcal{T}$ is not admissible; $\mathcal{T}_1$ is admissible.
- $\{(c, a)\}$ is an att-explanation for a.

Summary

Given an AA framework F and some non-admissible argument a in F, why isn't a admissible?

Two different notions of explanation and their computation

- Notions of explanation:
  - Explanation in argument view: *arg-explanation*
  - Explanation in attack view: *att-explanation*
- Explanation computation: prune dispute trees

Questions?