NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Interpretability

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University

Terminology

Understandability

- **Understandability** (or **equivalently**, intelligibility) refers to the characteristic of a model to make a human understand its function – how the model works – without any need for explaining its internal structure or the algorithmic means by which the model processes data internally

Comprehensibility

- **Comprehensibility:** when conceived for machine learning models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion

Interpretability

- **Interpretability:** it is defined as the ability to explain or to provide the meaning in understandable terms to a human.

Explainability

- **Explainability:** it is associated with the notion of explanation as an interface between humans and a decision maker
  - that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans

Explicability

- **Explicability:**
  - Making AI decisions obvious to a human being (i.e. a human being can understand the reason behind an AI decision without explanation)
  - Might not be the optimal solution!

Transparency

- **Transparency:** a model is considered to be transparent if by itself it is understandable. A model can feature different degrees of understandability.

Explainable AI through Argumentation

What is Argumentation?

- Evaluate “possible conclusions” by considering reasons for and against
  - Constructing pros and cons arguments
  - Evaluating arguments accordingly
- Resolve conflicts (within or across “agents”)
- Often studied and applied in
  - Disciplines: philosophy, logic, law, artificial intelligence, computer science, etc.
  - Applications: decision-making, dispute resolution, negotiation, security, bioinformatics, etc.

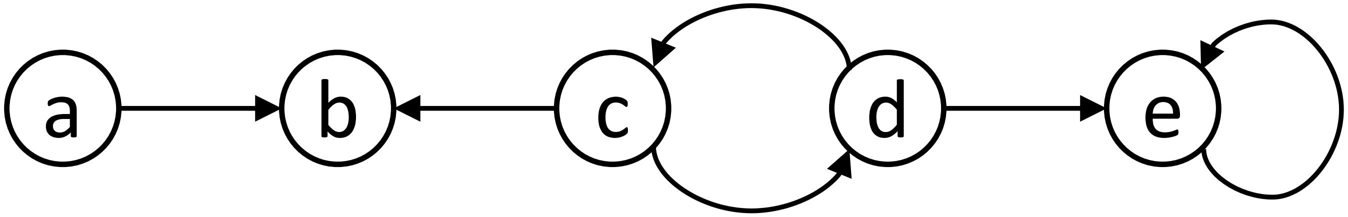
Argumentation: A Simple Example

- Abstract Argumentation
  - Arguments are “atomic”
  - Formalize relations (“attacks”) between arguments

- An **abstract argumentation framework** (AF) is a pair (A, R) where
  - A is a set of arguments
  - $R \subseteq A \times A$ is a relation representing “attacks”

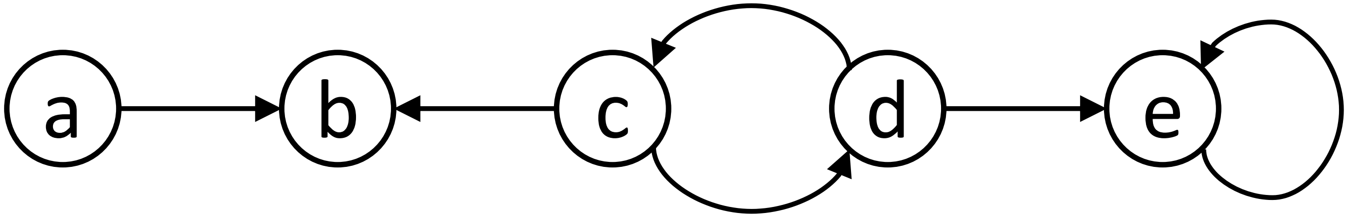
Argumentation: A Simple Example

- $A = \{a, b, c, d, e\}$
- $R = \{(a, b), (c, b), (c, d), (d, c), (d, e), (e, e)\}$



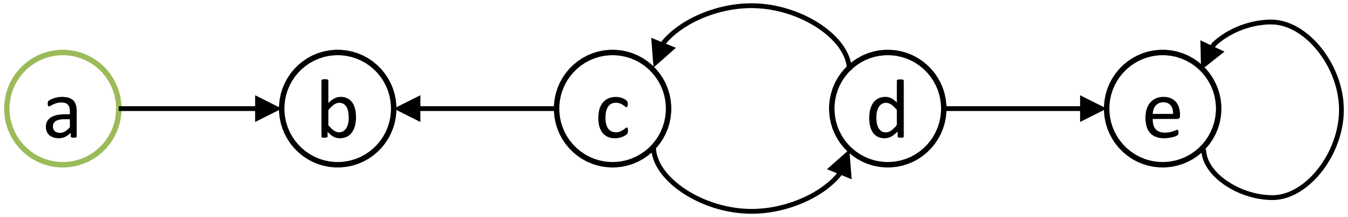
Argumentation: A Simple Example

- Conflict Free Set:
  - Given an AF $F = (A, R)$. A set $S \subseteq A$ is conflict-free (*cf*) in F, if, for each $a, b \in S, (a, b) \notin R$.



Argumentation: A Simple Example

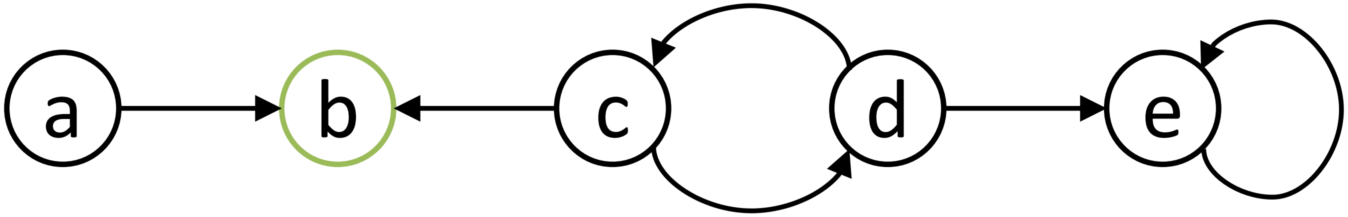
- Conflict Free Set:



❑ $cf(F) = \{\{a\},$

Argumentation: A Simple Example

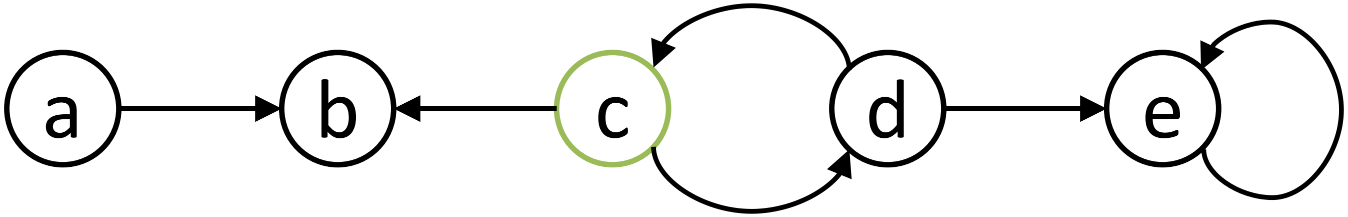
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}$

Argumentation: A Simple Example

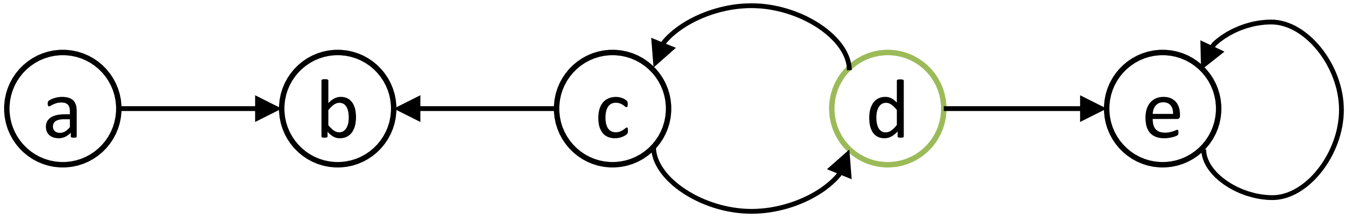
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}$

Argumentation: A Simple Example

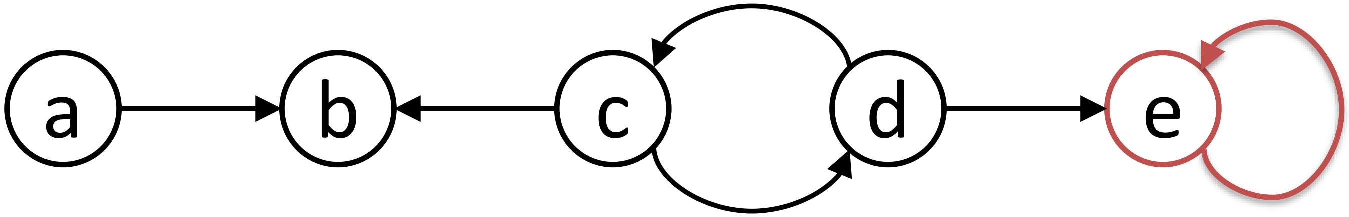
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}$

Argumentation: A Simple Example

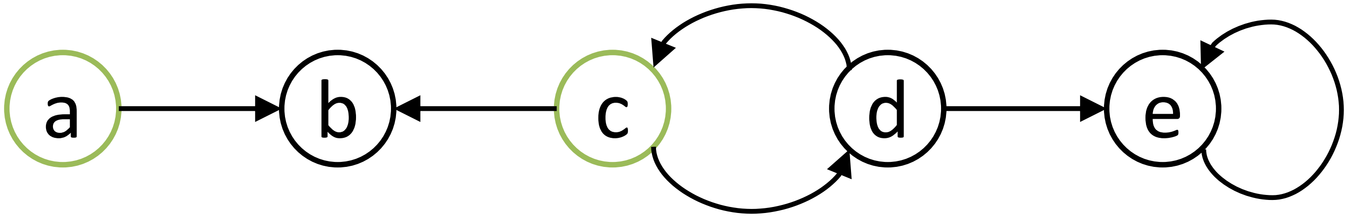
- Conflict Free Set:



❑ $cf(F) = \{\{a\},\{b\},\{c\},\{d\}$

Argumentation: A Simple Example

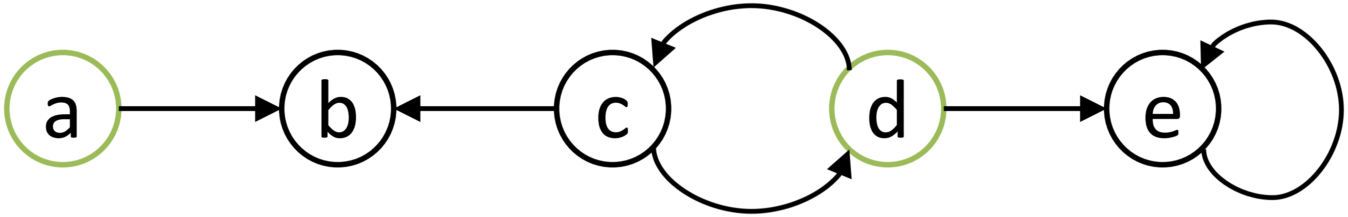
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}\}$

Argumentation: A Simple Example

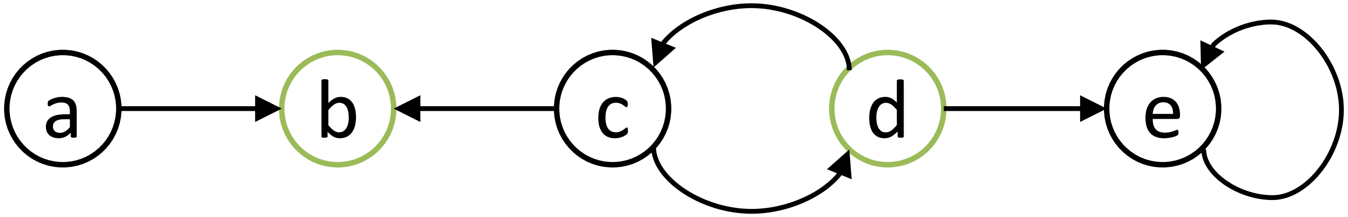
- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}$

Argumentation: A Simple Example

- Conflict Free Set:



❑ $cf(F) = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, c\}, \{a, d\}, \{b, d\}, \emptyset\}$

Interesting Reading

Alejandro Barredo Arrieta *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, vol. 58, pp. 82-115 (2020)

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Interpretability

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University