Explaining Identity-aware Graph Classifiers through the Language of Motifs

Alan Perotti
Paolo Bajardi
Francesco Bonchi
André Panisson

CENTAI

Presented at the International Joint Conference of Neural Networks - IJCNN2023

CENTAI

- Private company, working w/ universities
- Fundamental research
- Applied/industrial research projects
- Research projects (e.g. Horizon Europe)




Disclaimer:
In this talk, the focus is heavily on the XAI side.

Happy to put you in touch w/ colleagues working on:

- Graph Machine Learning
- (explainable) Link Prediction
- Social Networks (opinion dynamics etc.)
- Graph Counterfactuals
- Hypergraphs, higher-order data
- Complex systems
- Old-school graph algorithms (MST, search, etc.)

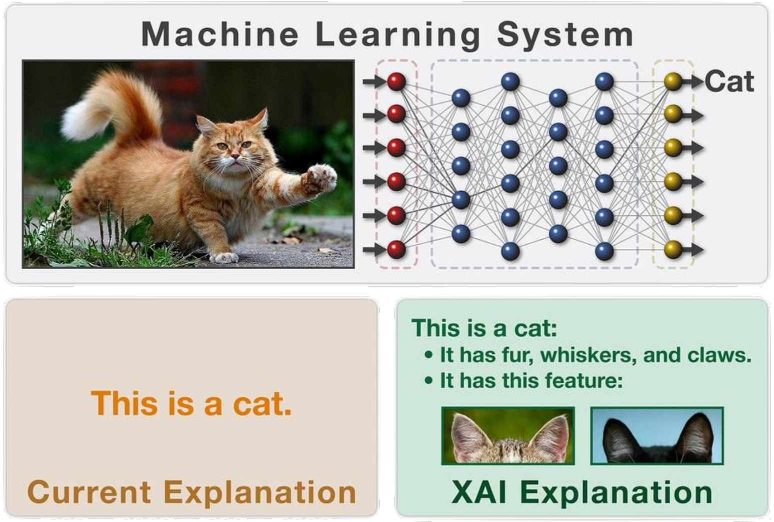
The Black-Box Problem

Many modern ML models are hard to interpret and it is difficult to understand why they make a certain decision or recommendation. This might cause several problems:

- No trust from experts.
- Biased systems.
- Right for the wrong reasons.
- GDPR non-compliance.
- Adversarial Vulnerability.

Explainable Artificial Intelligence

Intuition: decorate model's output with additional information.



Very hot area: much fundamental research to be done, strong interest from private companies, European research calls, etc.

The intuition behind this paper

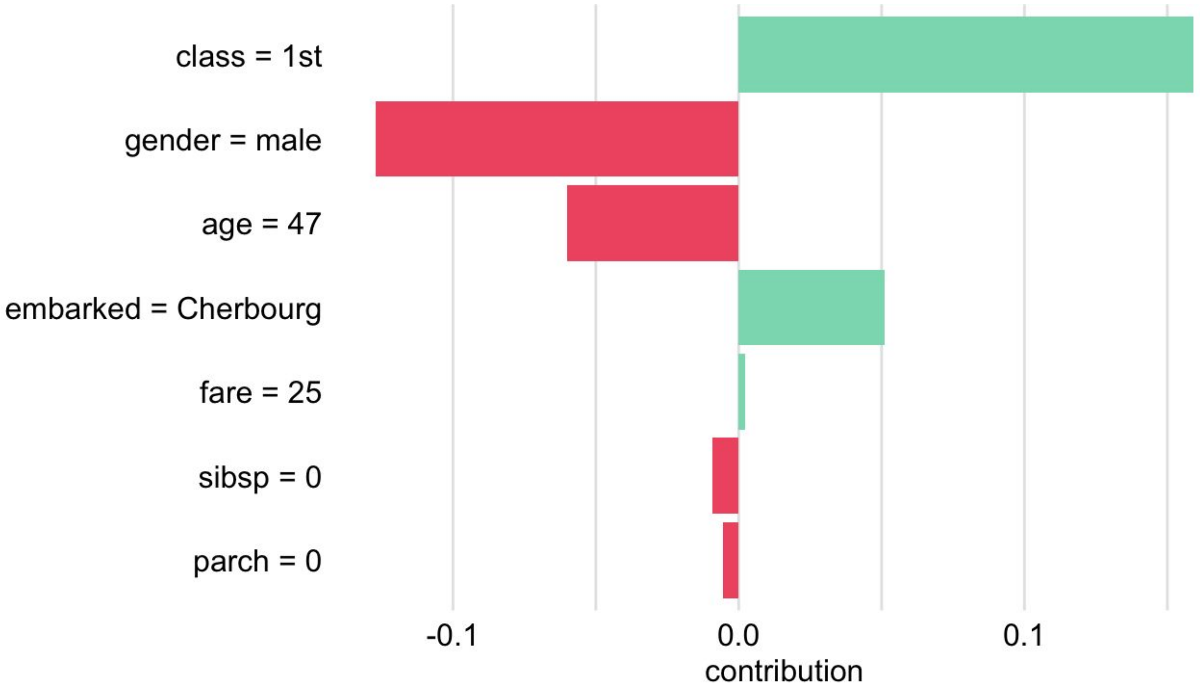
ML input language

VS

explanation language

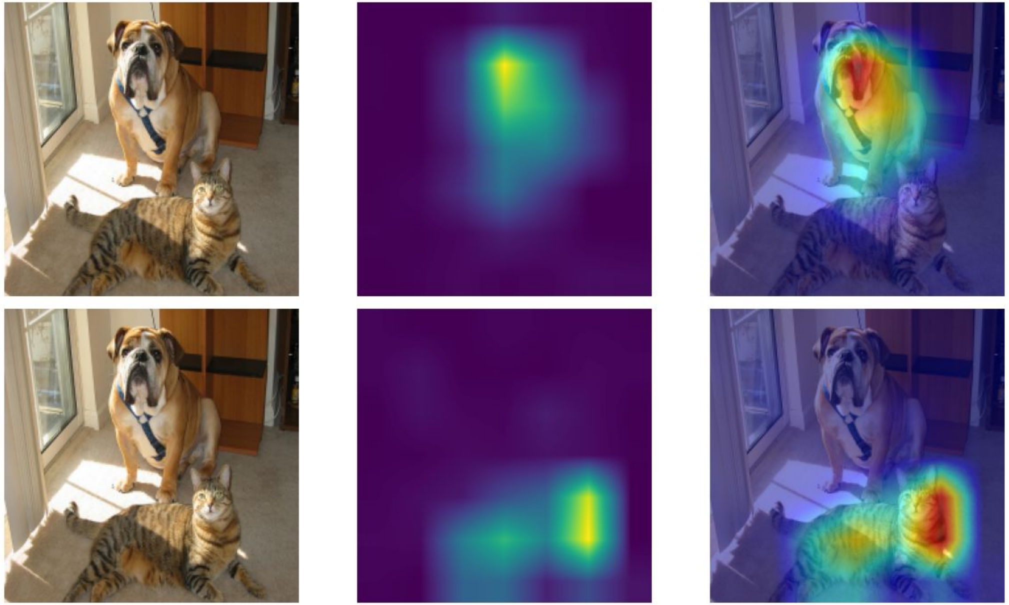
ML input language VS XAI explanation language

Attribution-based explanations on tabular data



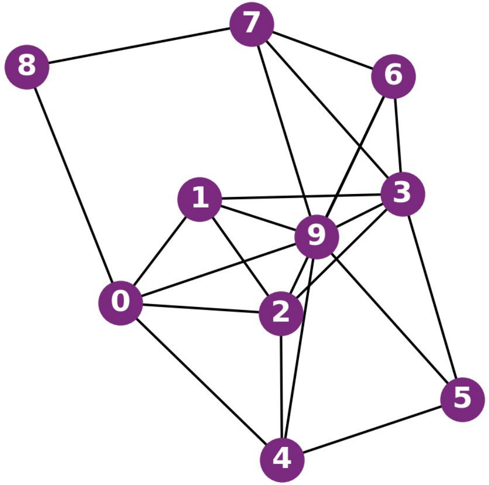
ML input language VS XAI explanation language

Heatmaps for Computer Vision ML models



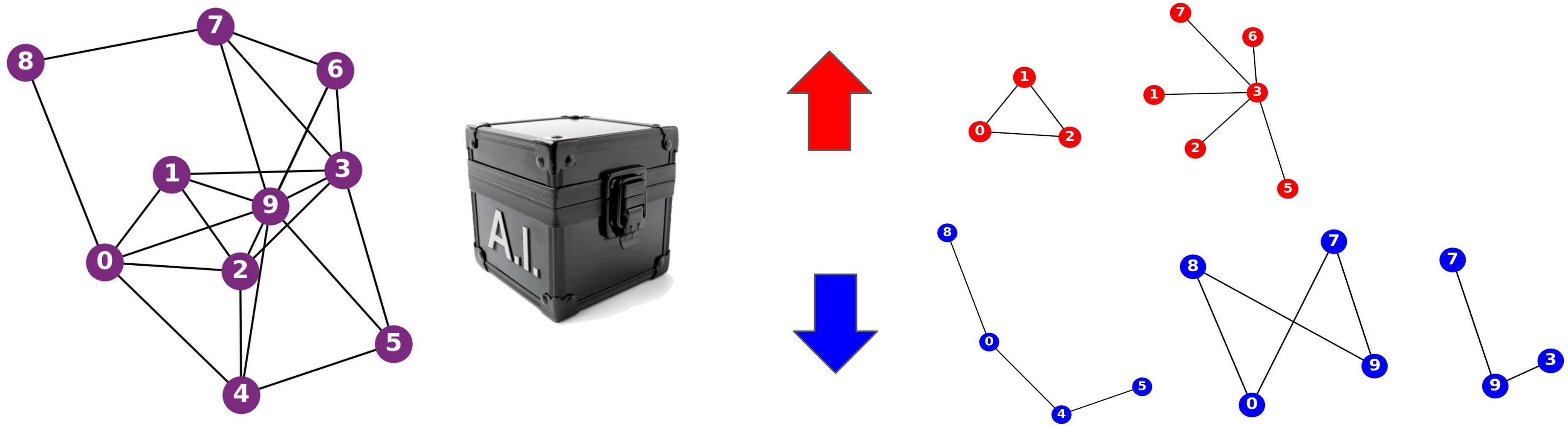
ML input language VS XAI explanation language

Graph classification

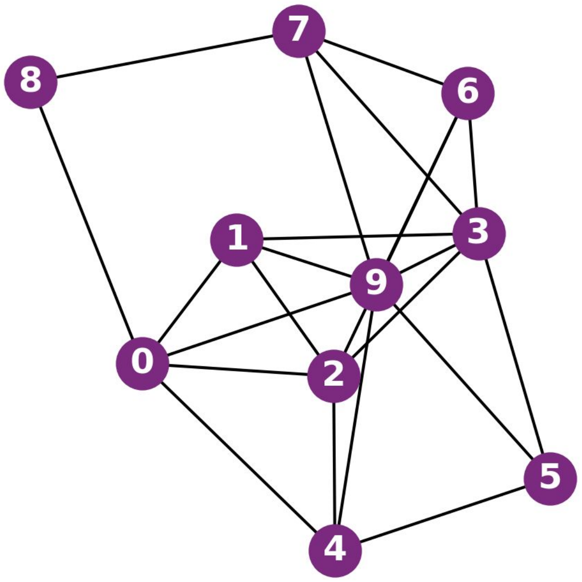


ML input language VS XAI explanation language

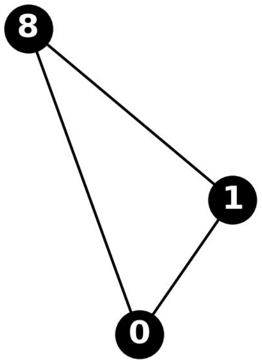
Graph classification and motifs (connected subgraphs)



Node Identity, graphs, and motifs



G



M

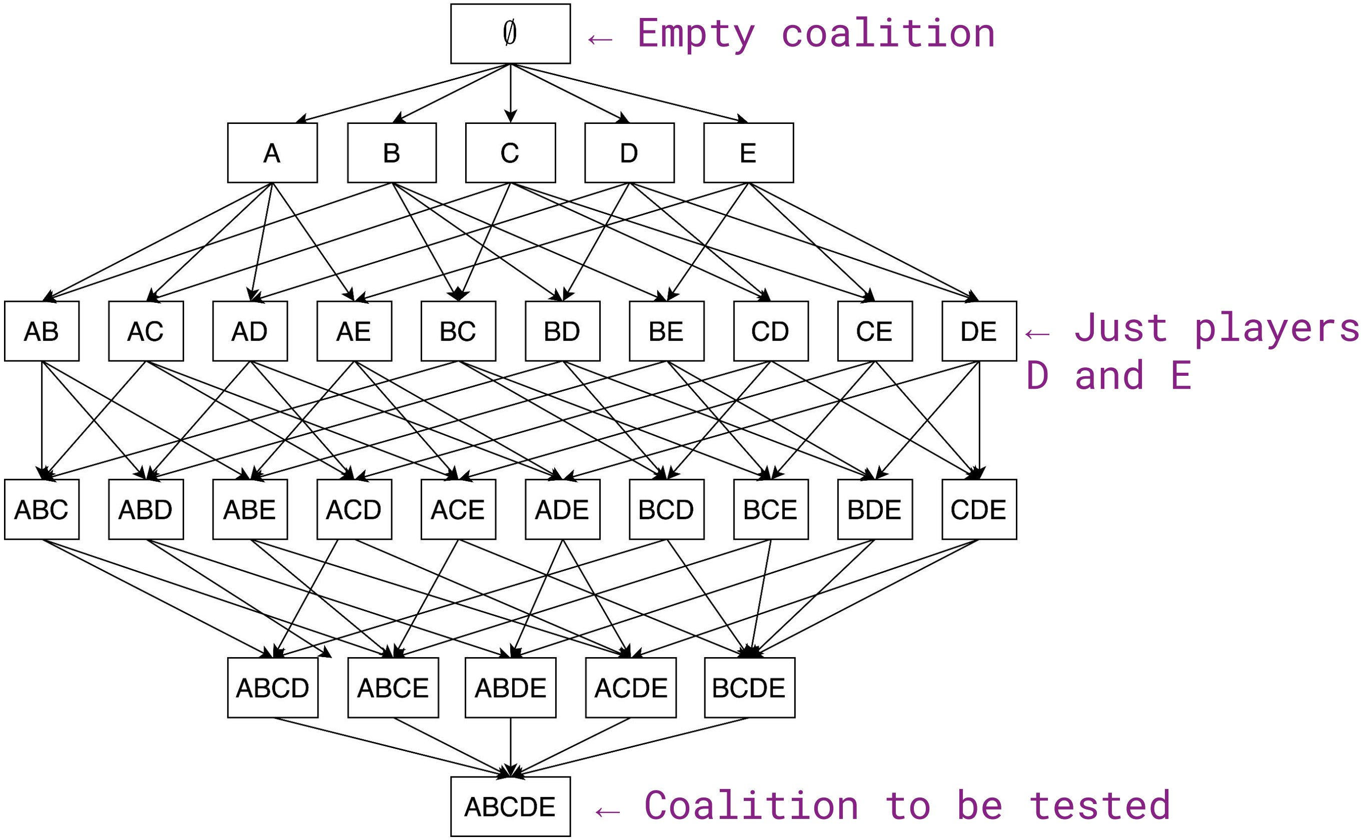
- Sub-graph of the induced complete graph, so might not occur.
- Links unique nodes.
- Connected, but this can be relaxed.

Problem Statement

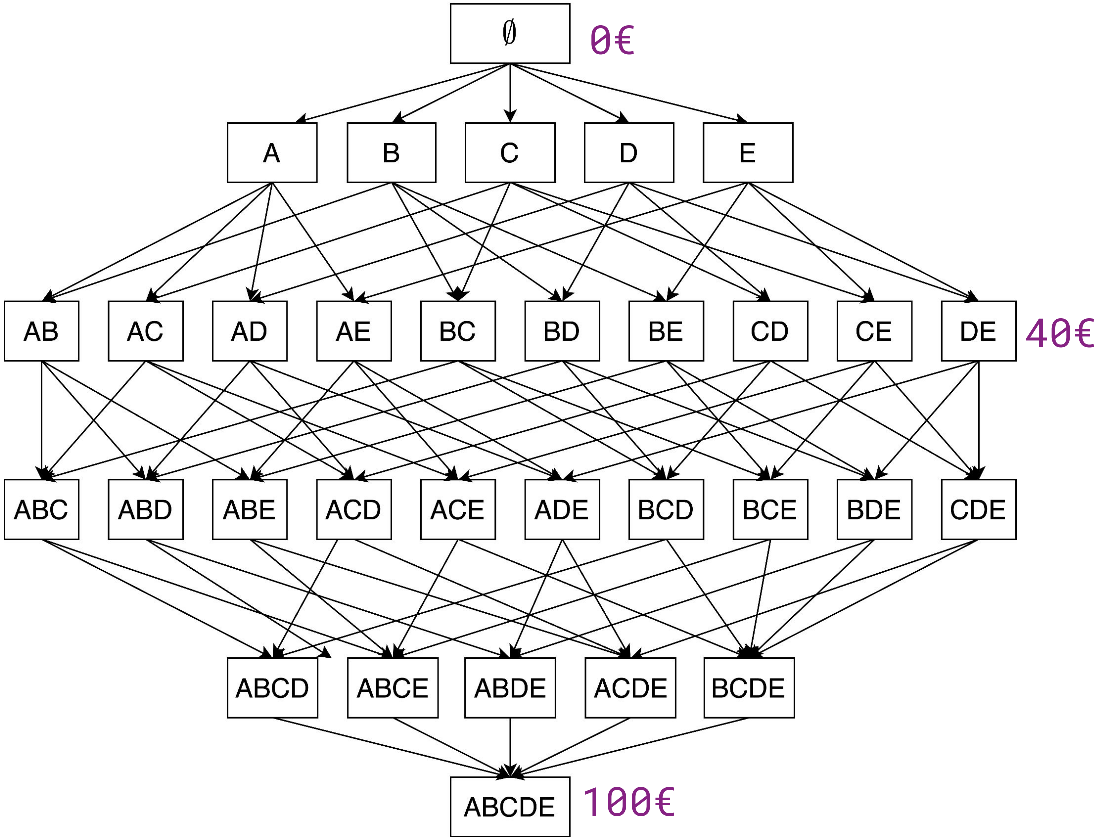
Given a graph $G \in \mathcal{G}$, a black-box $B : \mathcal{G} \to [0,1]$ and a set of motifs $\mathcal{M}$, the problem tackled in this paper is that of assigning an *explanation score* $\xi(G, B, M_i) \in [-1,1]$ to each motif $M_i \in \mathcal{M}$, quantifying the impact of the motif in explaining the label $B(G)$: a value close to -1 means that M_i is important in explaining $B(G) = 0$, a value close to 1 means that M_i is important for $B(G) = 1$.

Shapley value, 1951: a lattice of coalitions

Team ABCDE wins some money.
How to distribute among players?

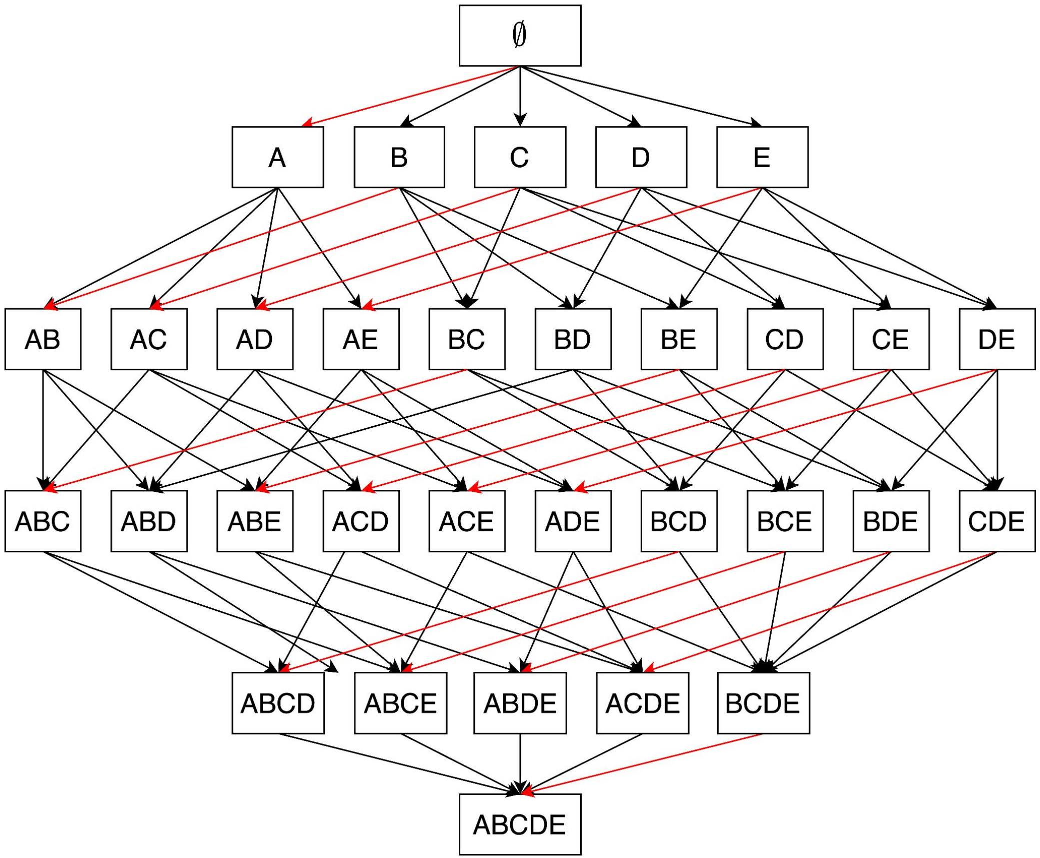


Shapley: every coalition has a value



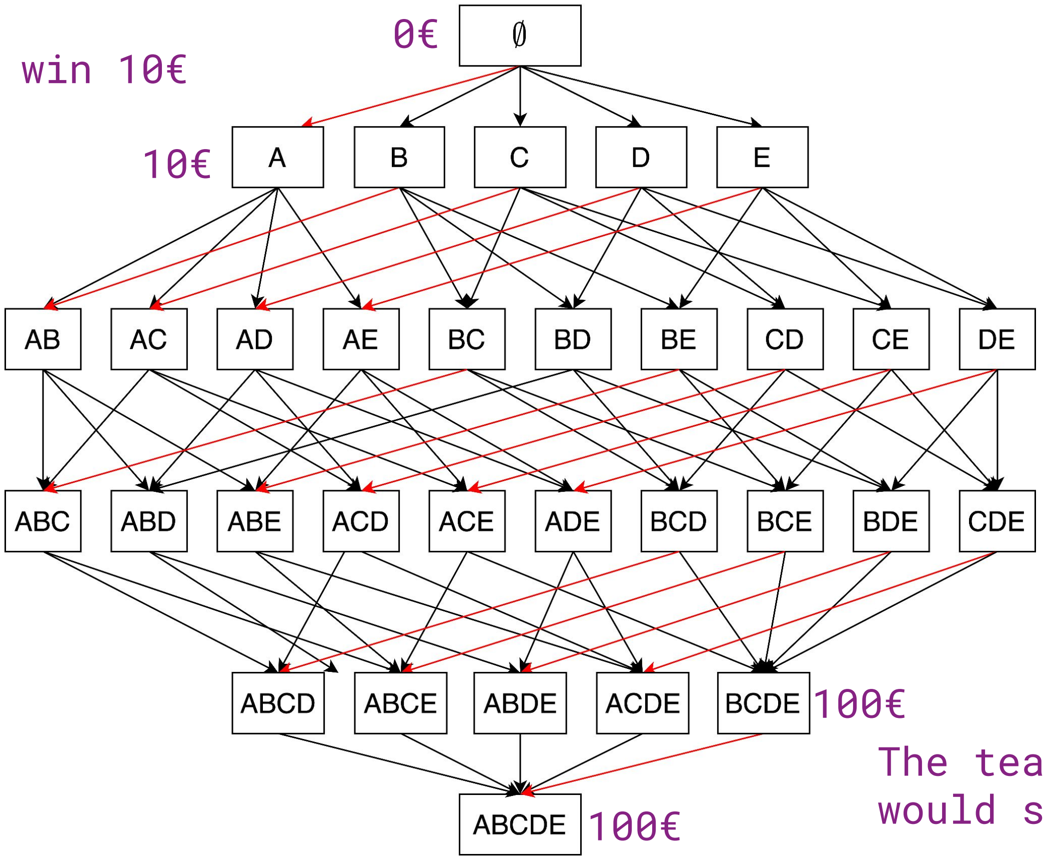
Shapley: every edge is adding a new player

Impact of player A



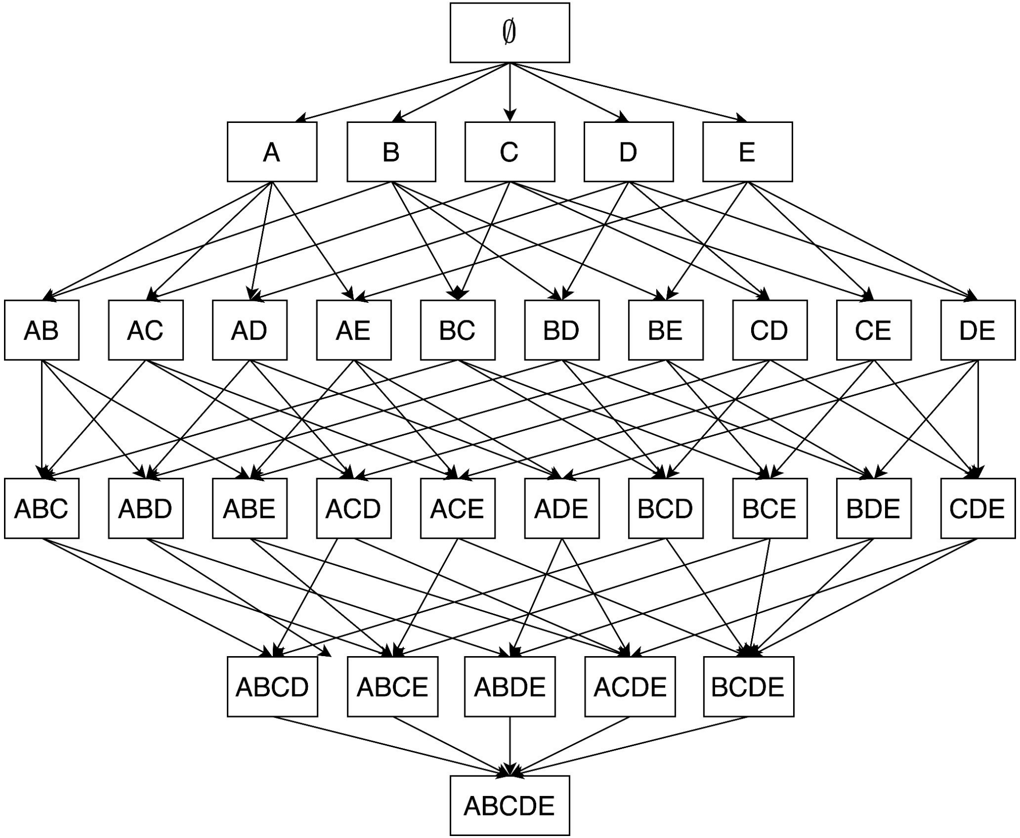
Shapley: every edge is adding a new player

A, alone, would win 10€



The team, without A, would still win 100€

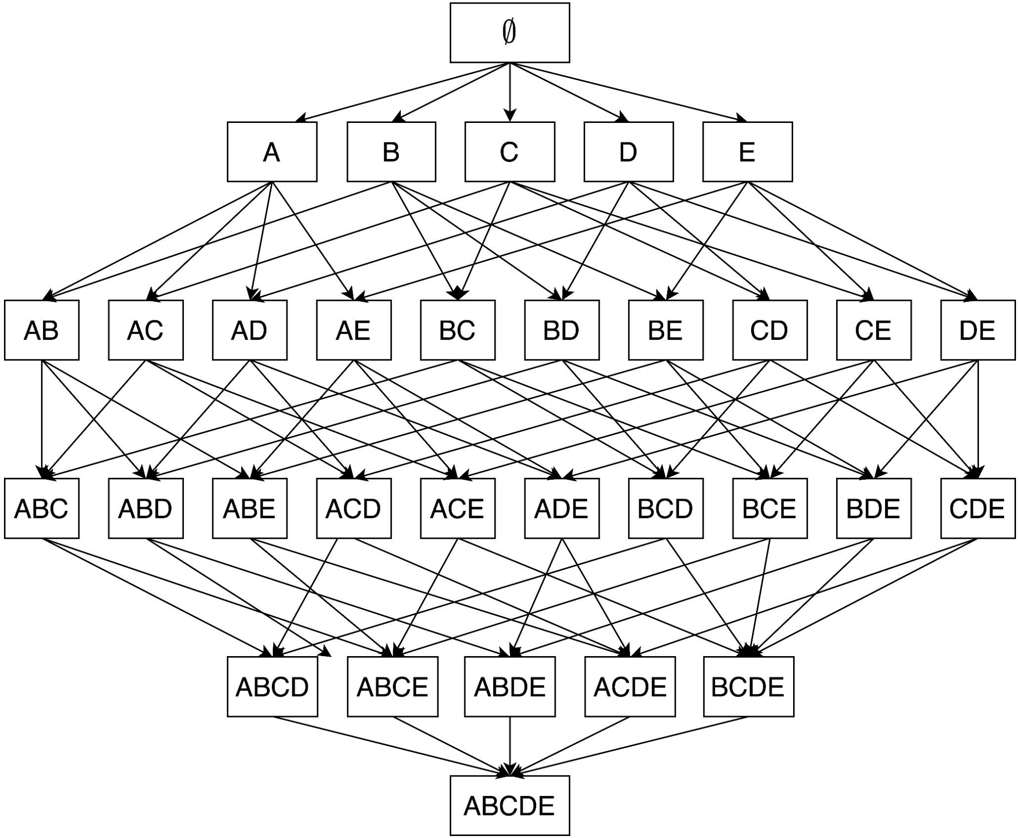
SHAP, 2017: same lattice, different interpretation



Players → Features
Coalitions/teams → Data Points
Team value → ML output (continuous)

Removing a player from the team and measuring the change in team value → masking a feature value in the data point and measuring the change in ML output

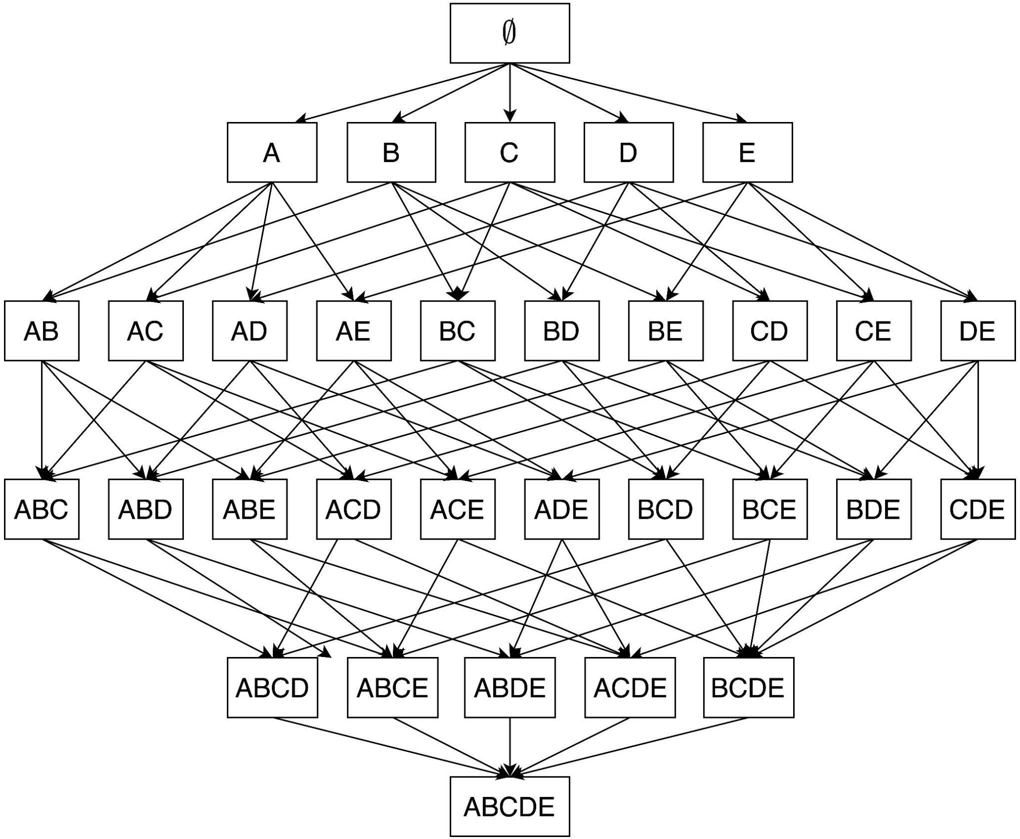
SHAP, 2017: same lattice, different interpretation



Pro: Arguably, XAI SOTA

Con: TWO devils in the details: scalability/approximation, feature removal

GraphSHAP, 2023: same lattice, different interpretation

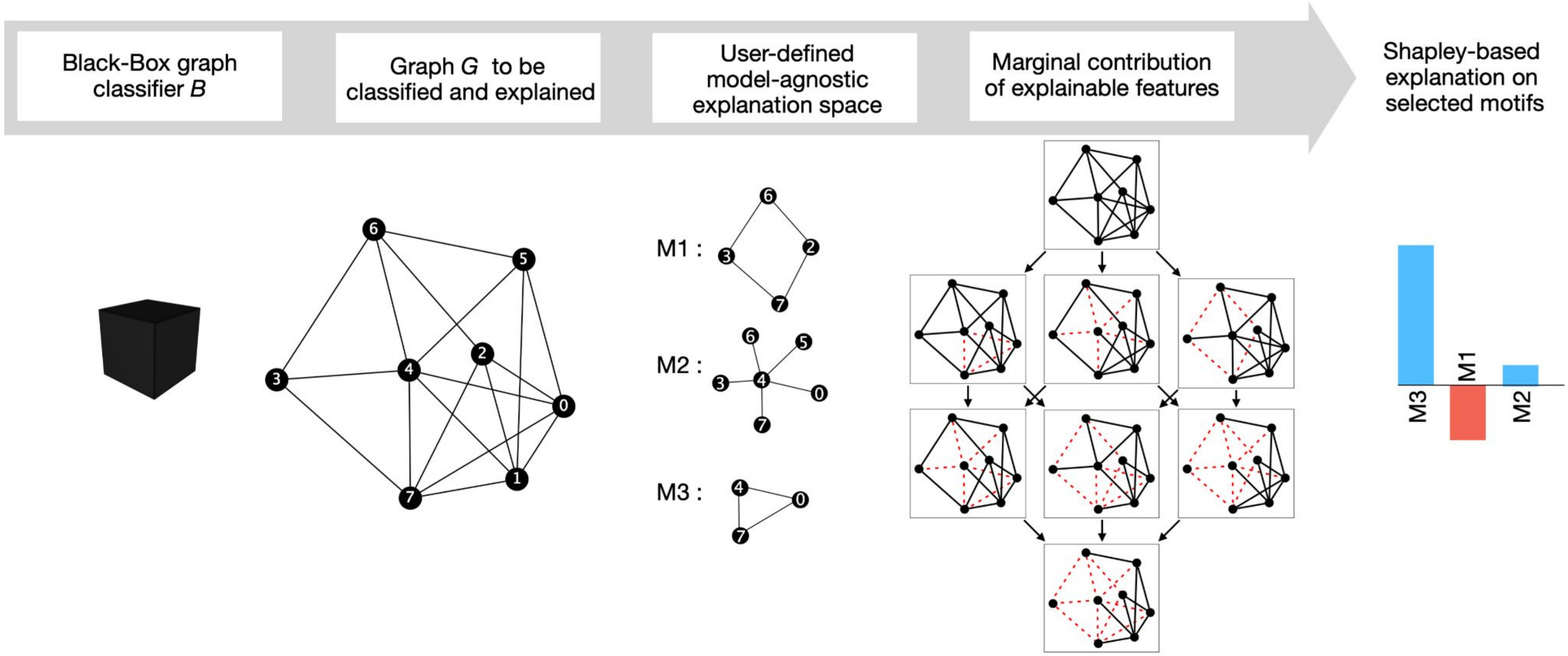


Players → MOTIFS
Coalitions/teams → Data Points
Team value → ML output (continuous)

Removing a player from the team and measuring the change in team value → masking a motif in the graph data point and measuring the change in ML output

(we inherit the same weaknesses)

GraphSHAP pipeline

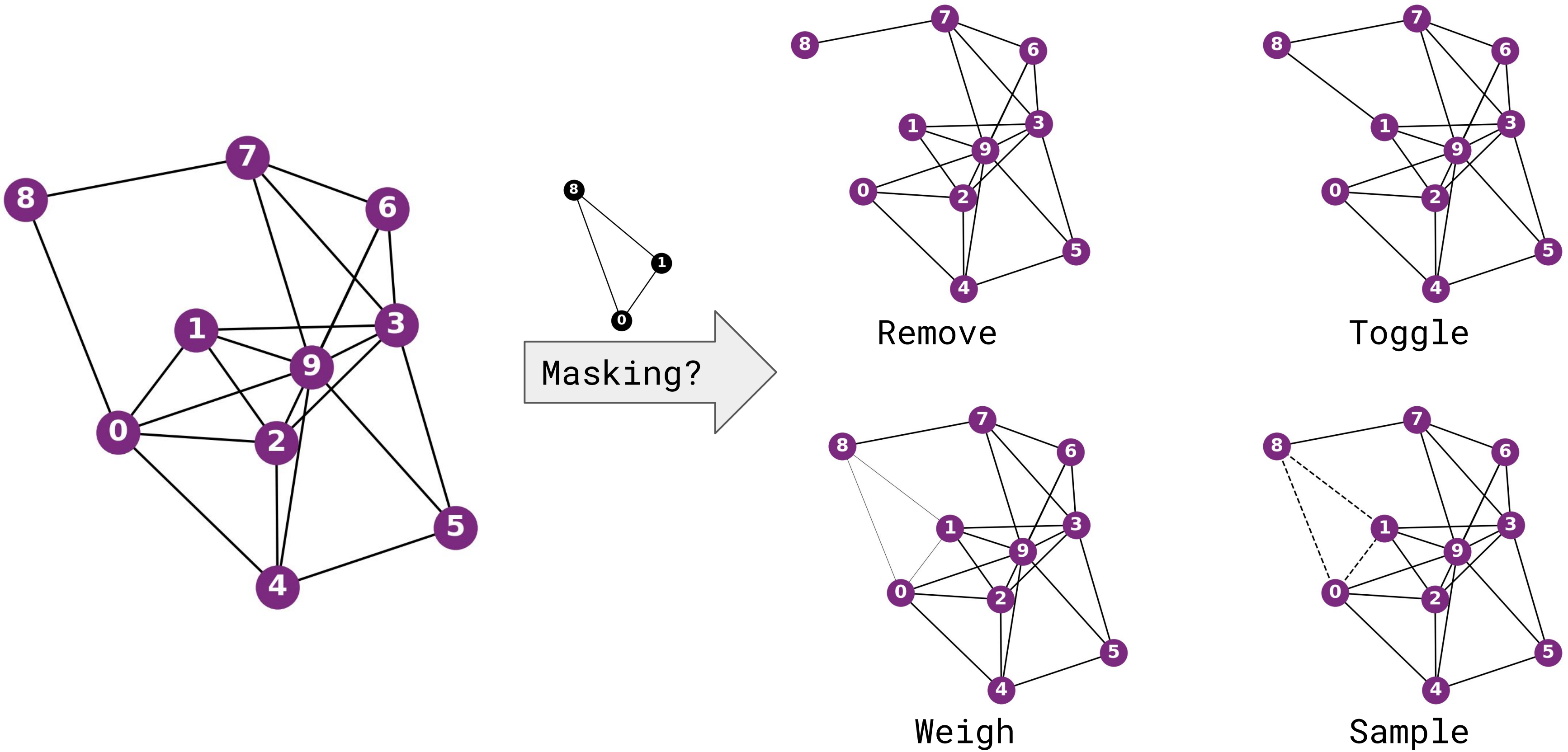


Shapley's heavy heritage 1/2

How do we translate the concept of *removing a player* in our graph-ML setting?

SHAP introduces the concept of *background dataset*, and copies values from other data points

Motif masking (with node identity)



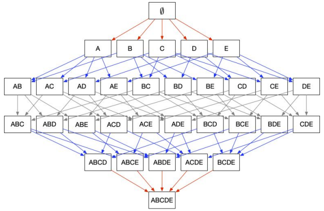
Shapley’s heavy heritage 2/2

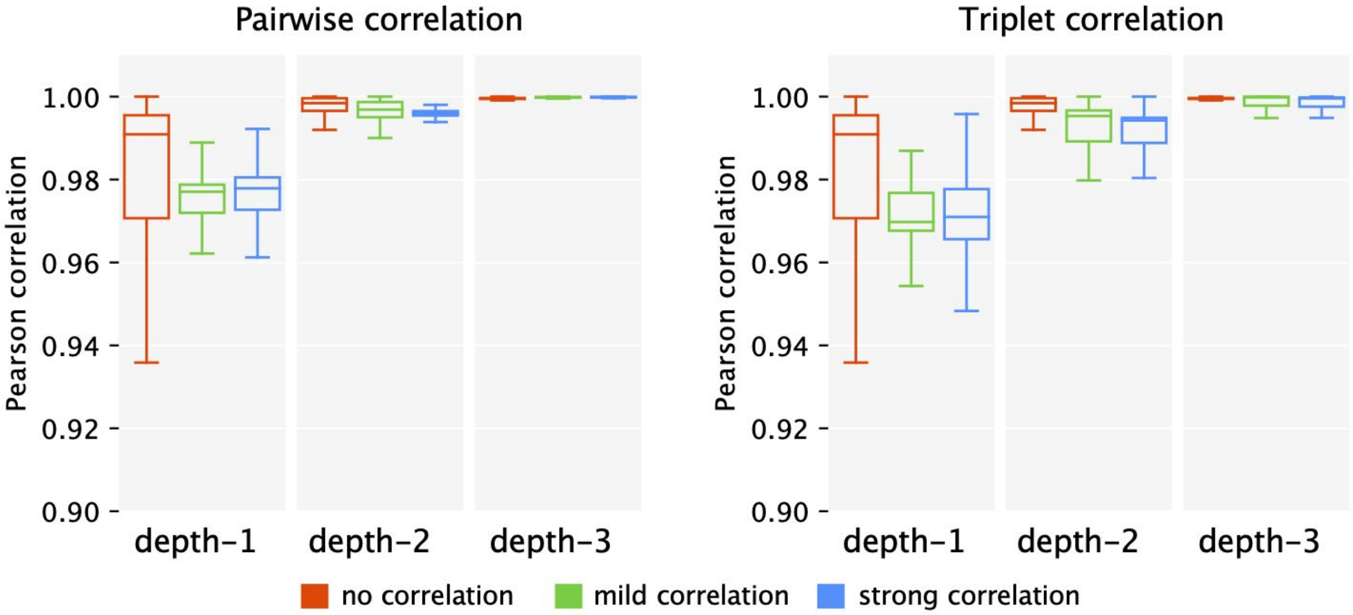
How do we deal with the lattice’s exponential computational complexity (wrt the number of features)?

SHAP introduces the concept of *budget*, and samples the budget according to heavily engineered heuristics

Experimental approximation

We found a strong approximation (with respect to the full Shapley lattice) in the first Shapley layer.





ABIDE dataset

Welcome to the Autism Brain Imaging Data Exchange!

ABIDE
Autism Brain Imaging
Data Exchange

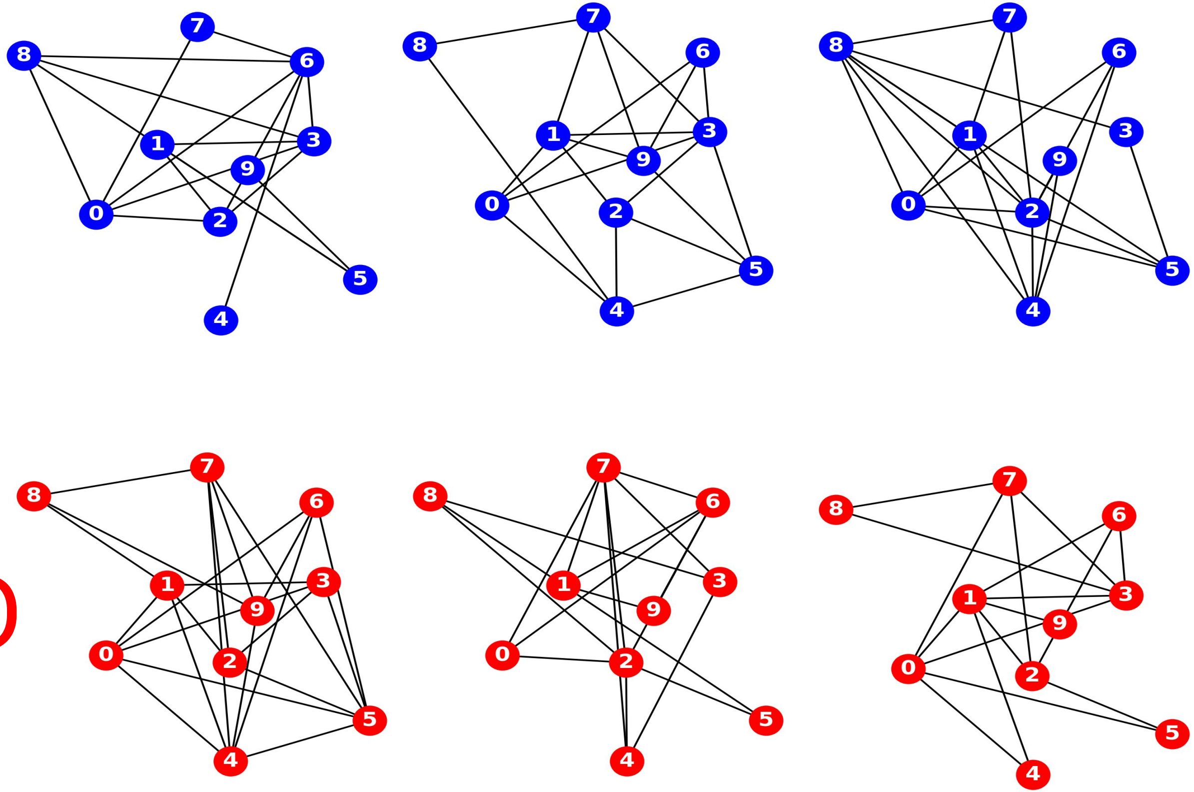
Introduction

Autism spectrum disorder (ASD) is characterized by qualitative impairment in social reciprocity, and by repetitive, restricted, and stereotyped behaviors/interests. Previously considered rare, ASD is now recognized to occur in more than 1% of children. Despite continuing research advances, their pace and clinical impact have not kept up with the urgency to identify ways of determining the diagnosis at earlier ages, selecting optimal treatments, and predicting outcomes. For the most part this is due to the complexity and heterogeneity of ASD. To face these challenges, large-scale samples are essential, but single laboratories cannot obtain sufficiently large datasets to reveal the brain mechanisms underlying ASD. In response, the Autism Brain Imaging Data Exchange (ABIDE) initiative has aggregated functional and structural brain imaging data collected from laboratories around the world to accelerate our understanding of the neural bases of autism. With the ultimate goal of facilitating discovery science and comparisons across samples, the ABIDE initiative now includes two large-scale collections: ABIDE I and ABIDE II. Each collection was created through the aggregation of datasets independently collected across more than 24 international brain imaging laboratories and are being made available to investigators throughout the world, consistent with open science principles, such as those at the core of the International Neuroimaging Data-sharing Initiative. For details about these initiatives visit the collection specific pages: ABIDE I and ABIDE II.

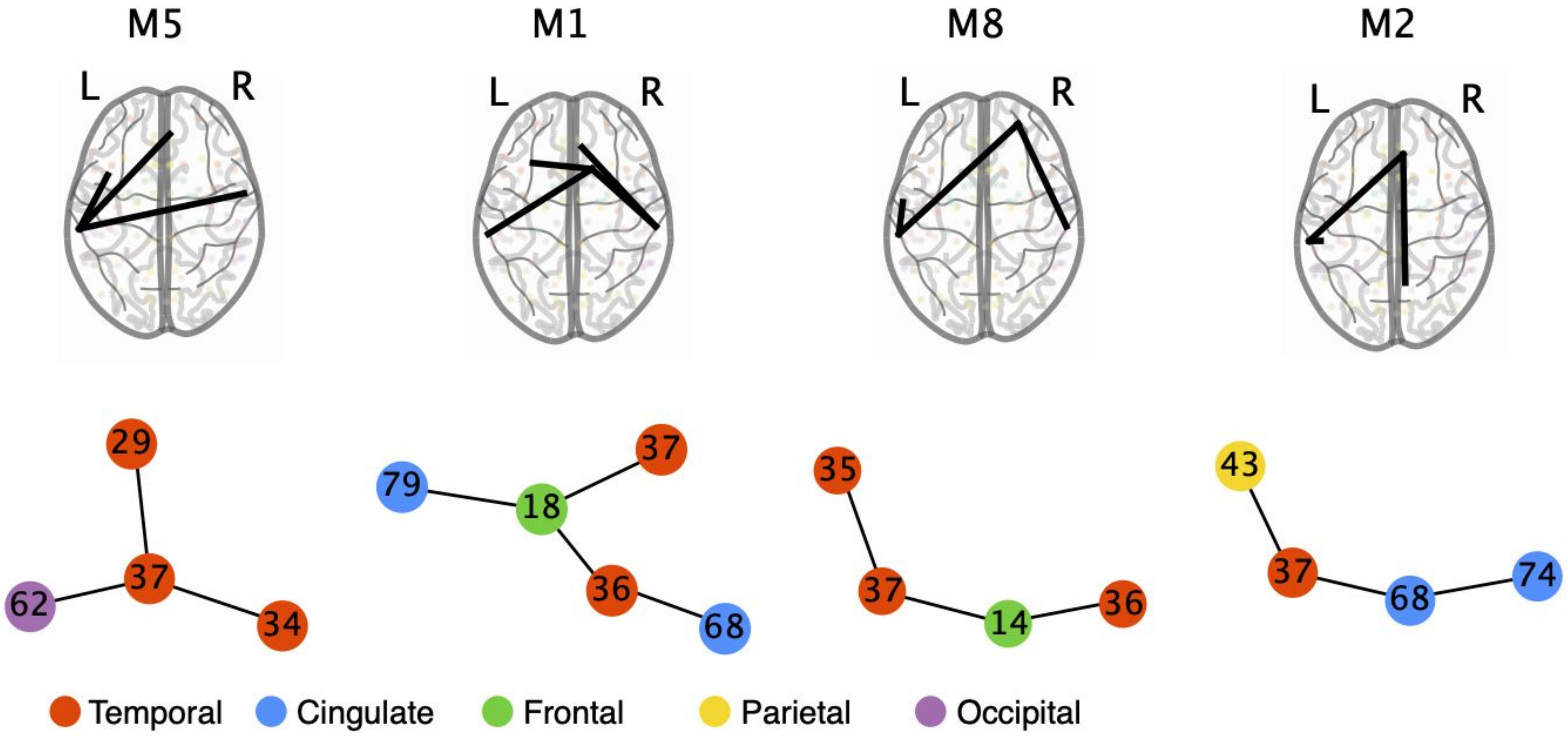
Graph Classification

TD

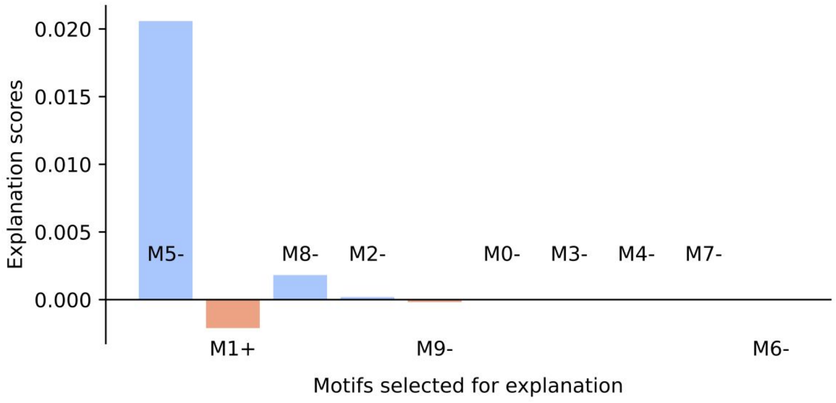
ASD



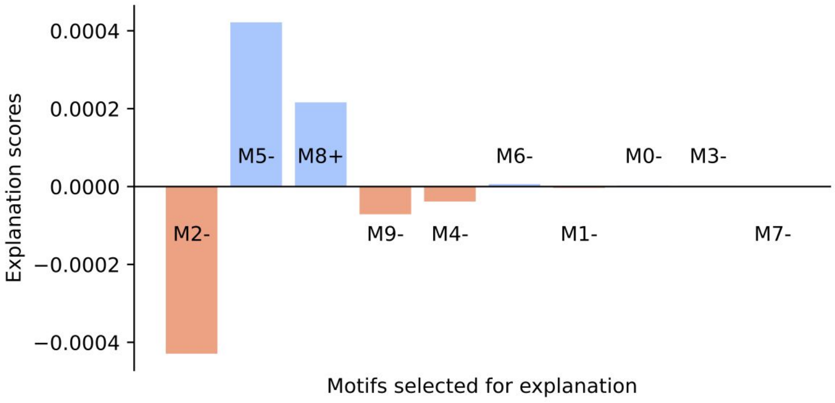
ABIDE motifs



Local explanations

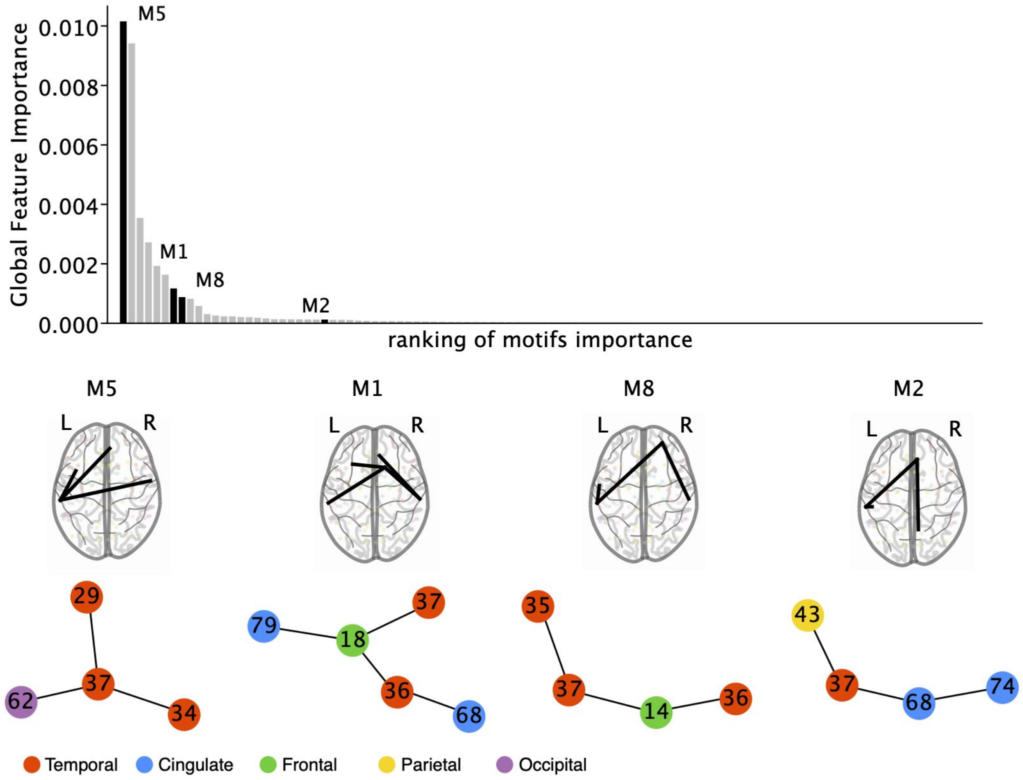


Patient A (ASD)



Patient B (ASD)

Global explanations



Take-home message

We developed a Shapley-based XAI algorithm for graph classification w/ node identity.

GraphSHAP computes attribution scores (a.k.a. feature importances) for a set of arbitrarily-defined motifs.

Conclusions: pros and cons

PROs:

- Custom, high-level explanation language
- Scalable algorithm
- Rooted in Shapley's game theory

CONs:

- Requires node identity (so far)
- Requires motifs
- Masking is arbitrary




Thanks!

Perotti, Bajardi, Bonchi, Panisson, “Explaining Identity-aware Graph Classifiers through the Language of Motifs”. International Joint Conference on Neural Networks 2023