

Explainable Query Answering (XQA): Going beyond traditional explanation approaches

Pouya Khani
Aarhus University

eXplainable Query Answering (XQA)

- We need explainability and transparency in data analysis systems.
- AI and ML is not the only component of a data analysis system.
- Need for **query answering explanations**, especially for aggregate queries.
- **Aggregate queries** summarize large and complex datasets into actionable insights.
- **Limitations of traditional explanation methods:**
 - ✓ Rely on purely statistical associations (may overlook critical causal dependencies).
 - ✓ Require fully specified causal graphs (usually infeasible due to complexity, domain uncertainty, and computational cost).
- **My PhD project:**
 - ✓ Focus on eXplainable Query Answering (XQA).
 - ✓ Going beyond traditional Explainable AI (XAI) methods by incorporating partial causal knowledge into non-causal approaches.
 - ✓ develop **partial causal explanations** for static and temporal data, and queries with and without join (single- or multi-table).
 - ✓ Allows analysts to benefit from **causally-informed interpretations without needing fully specified causal models**.
 - ✓ Study the **tradeoff between accuracy and complexity** of explanations.

Aggregate Query Explanation

- **Aggregate query results** alone lack explanation for observed outcomes.
- **Example.** consider a developer survey data:

```
SELECT AVG(Salary)
FROM data
WHERE predicate p
```

Numeric results alone do not reveal importance of predicates to the final result, like {Education = "PhD"} or {Age = "under 18" AND Major = "Computer Science"}.

- ✓ Explanations can highlight the impact of individual or combined predicates on the aggregate outcomes.

Why Do We Need Causal Explanations?

Non-causal explanations for aggregate queries can be generated through statistical correlation or game-theoretic attribution, but these methods often misrepresent the true importance of data features due to ignoring causal relationships.

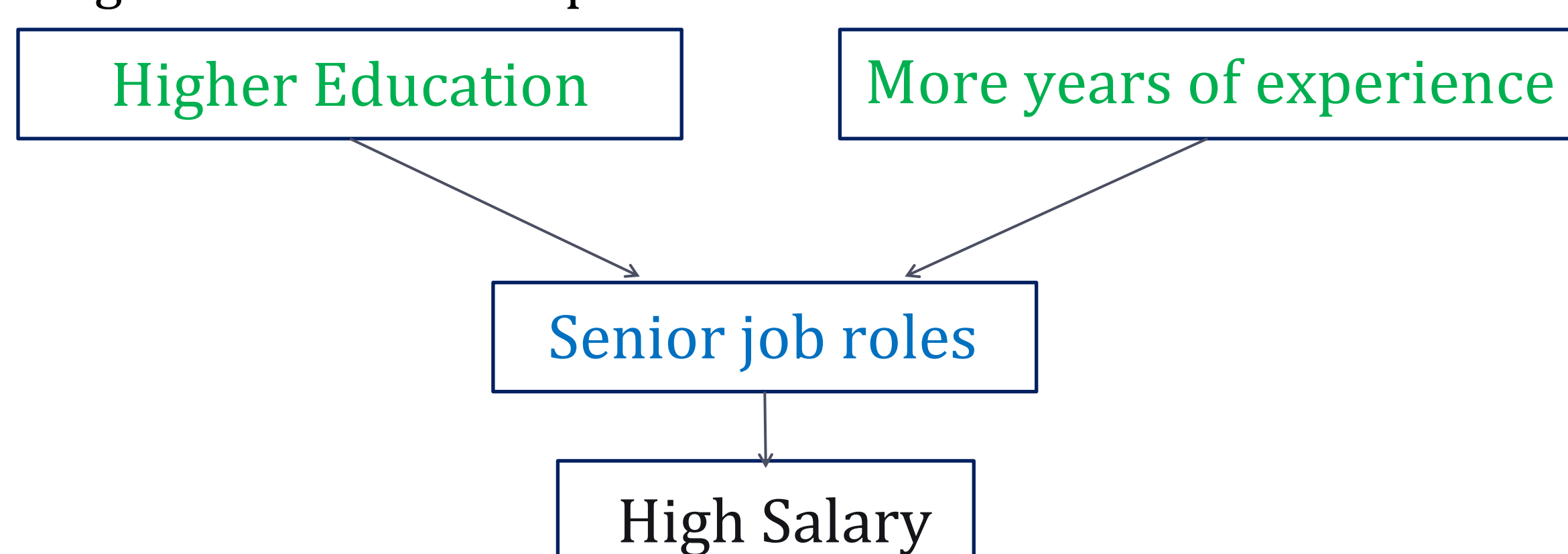


Figure 1: a naive explanation might attribute high salaries directly to **senior job roles** (non-causal explanation) without considering that **higher education** or **greater experience** causally precedes such roles (causal explanation).

[1] Stack Overflow, Developer Survey, 2021. URL: <https://survey.stackoverflow.co/2021>.

Causal Banzhaf Value for aggregate query explanations

Given a set of features $N = \{1, 2, \dots, N\}$, a target variable Y , and a partial causal Directed Acyclic Graph (DAG) G , CBV computes the importance of a specific predicate $p(i = u)$ (feature i with value u) as follows:

$$\beta_{p(i=u)}^C = \frac{1}{2^{n-1}} \sum_{S \in \mathcal{S}_i} \sum_{r \in \mathcal{R}(S)} [v(S_r \cup \{p(i=u)\}) - v(S_r)]$$

Where $\mathcal{S}_i = \{S \mid A_i \subseteq S, S \subseteq N \setminus \{i, Y\}\}$ defines the set of **causally valid subsets**, ensuring all causal ancestors A_i of feature i are included. r represents a specific realization (value assignment) of the features in the subset S . $V(S_r)$ is the expected value of Y conditioned on subset S taking realization r .

- CBV offers **significant computational efficiency over BV**:
- By focusing only on causally valid subsets.
- Especially notable for attributes with many causal preceding relationships.

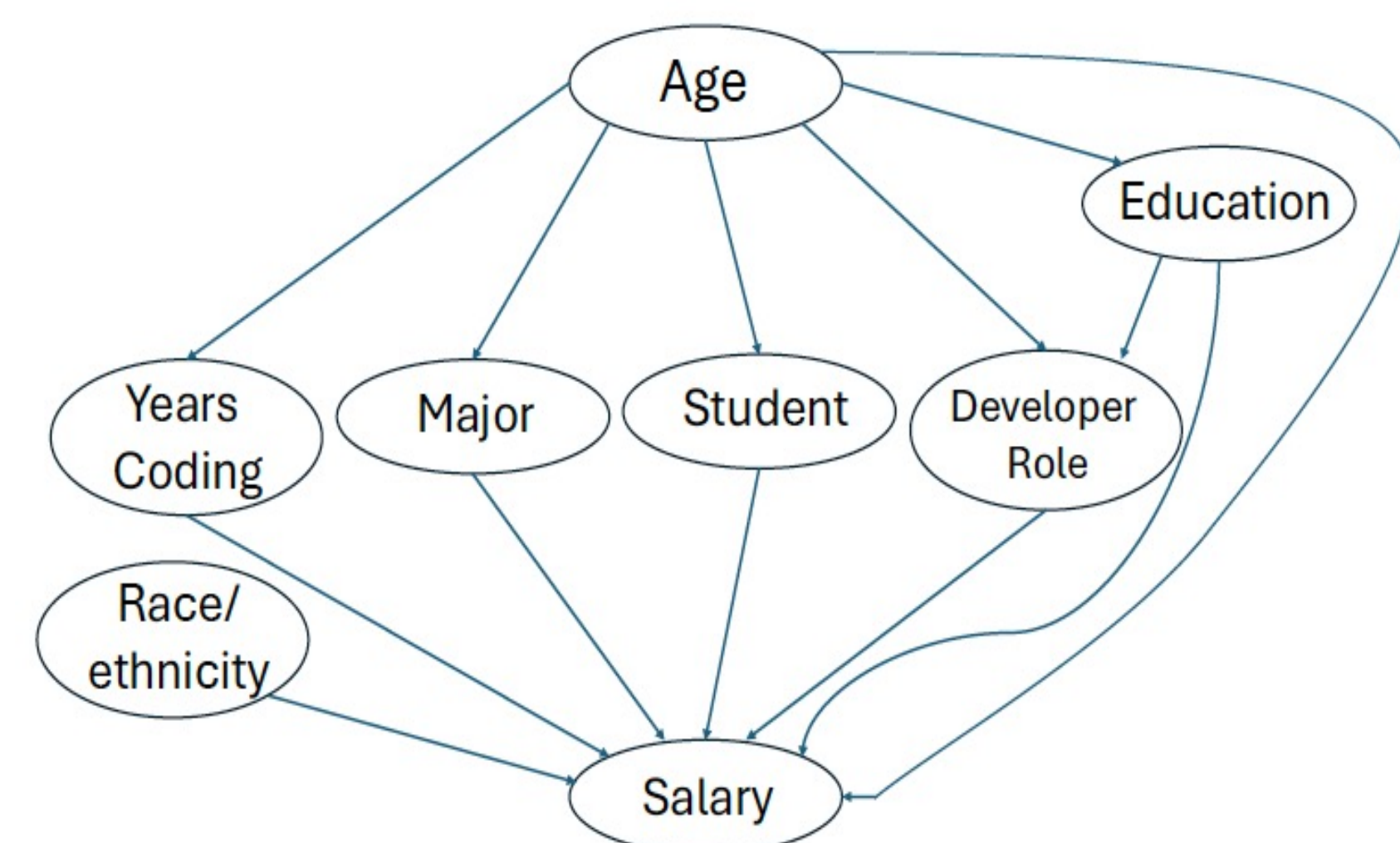


Figure 2: Partial Causal DAG, adapted from the full causal DAG from domain knowledge.

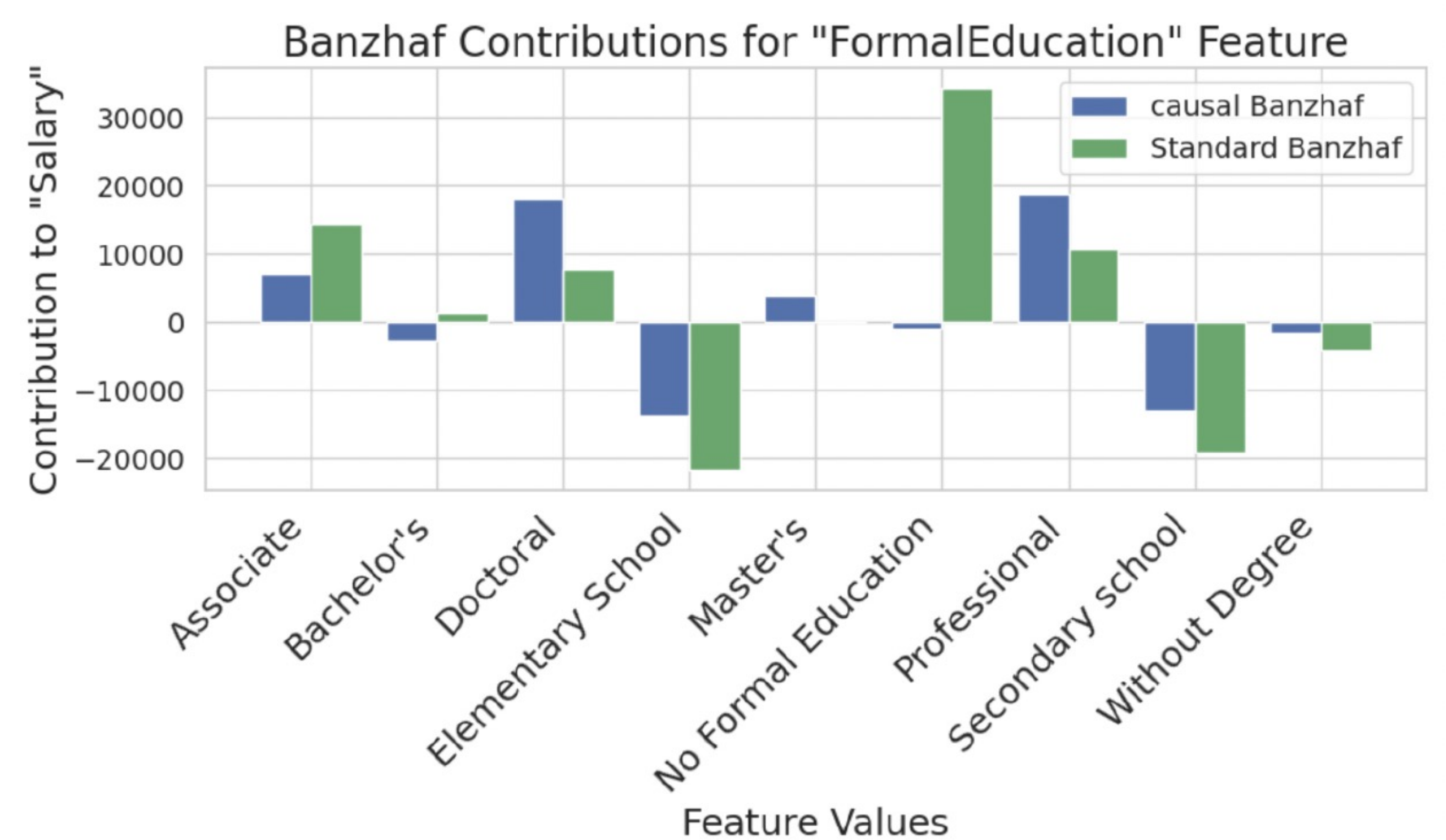


Figure 3: CBV and BV stack overflow [1] survey predicates contribution to salary (as an example, we just put contributions of one of attributes in this poster). CBV obtains predicate importance in line with the partial causal DAG in figure 3. This figure illustrates contributions of all values for "Formal education level" to "Salary" according to BV and CBV.

Future works

- ✓ Extend CBV to the **extended version**:
 - Calculate contribution of both individual and **combined predicates**.
 - Utilize and develop **optimizations and approximations** to increase speed.
- ✓ Causal Banzhaf Value for **temporal data and queries**, with dynamic causal dependencies and moving average query.
- ✓ Causal Banzhaf Value for **join-aware explanations**, suitable for multi-table data with inter-connected causal dependences.

✓ Do you like to collaborate or discuss more? Just send me an email:
pouya.khani@cs.au.dk