

Peer Effects and Educational Inequalities in Higher Education Choices

A Hybrid AI and Econometric Approach Using Knowledge Graphs

Host institutions: Sorbonne Université (LIP6) and Université Paris-Panthéon-Assas (LEMMA)

Disciplines: Artificial Intelligence, Computer Science, Economics.

Keywords: Peer effects; Knowledge graphs; XAI; Education economics; Causal inference; Parcoursup.

Supervisors: Christophe Marsala (Computer Science, LIP6) [lead supervisor]; Maria Rifqi (Computer Science, LEMMA and LIP6); Sandra Cavaco (Economics, LEMMA)

1. Context and Motivation

Each year, more than 800,000 French students apply to higher education through Parcoursup, selecting from over 20,000 programmes. Even after controlling for academic performance, choices vary markedly by social background, gender, and geographic origin — patterns that standard rational-choice models fail to fully explain. A growing body of literature points to **peer effects** as a key driver: the social, academic, and gender composition of a student’s class or school significantly influences aspirations and applications [4, 5]. Yet their impact on **educational choices** remains largely understudied [2, 7].

The **Parcoursup database** (PSupDB), with data since 2018 accessible via secure CASD access, is currently fragmented across more than 700 tables and is largely underused by the research community. It contains rich information on students (academic records, social origin, geography), schools, programmes, and the full set of application choices made by each student. Exploiting its relational structure requires advanced computational tools that do not sacrifice interpretability for social scientists.

The **central question** addressed in this thesis is: to what extent does the social, academic, and gender composition of a student’s peer group influence their higher education application choices, and how can these peer effects be modelled in a way that is simultaneously *causally credible*, *computationally scalable*, and *interpretable* for social scientists and policy makers? This question gives rise to four interrelated areas of research that cover both Computer Science (CS) and economics:

1. **Knowledge representation and relational modelling (CS).** How can the complex, multi-relational structure of PSupDB be faithfully encoded in a Knowledge Graph (KG)? What ontological choices best capture peer relationships, application portfolios, and institutional contexts? How can the KG be enriched with external RDF sources without redesigning the schema?
2. **Interpretable and neuro-symbolic rule learning (CS).** How can symbolic rules be mined from the KG in a way concise, non-redundant, and meaningful to social scientists? How can neuro-symbolic approaches, combining the predictive power of graph embeddings with the transparency of logical reasoning, be designed for structured educational data? How can contrastive and counterfactual explanations be generated from heterogeneous models (symbolic, neural, econometric) to support interpretability and validation by domain experts?
3. **Quantification and structure of disparities (Economics).** How the higher education choices across schools and classrooms are attributable to peer composition (education, social origin, gender)?
4. **Heterogeneity, mechanisms, and policy-relevant counterfactuals (Economics).** Which students are most sensitive to peer influence and through which reference subgroups (top performers, same-gender peers, students of similar social background)? What would happen to inequalities in application choices under alternative school or class compositions — through zoning reform, academic tracking, or ability grouping?

The **originality** of this thesis lies precisely in the articulation of axes 1–2 and axes 3–4: symbolic rules and KG-based patterns discovered computationally will guide econometric model specification, while causal estimates will in turn constrain and validate the learned representations.

2. Scientific Contributions

2.1. Computer Science Axis: Knowledge Graphs and Neuro-Symbolic Approaches

Building an Educational Knowledge Graph. The core technical contribution of the thesis is the design and implementation of a KG encoding the full PSupDB as RDF triples. The ontology will include core classes and relational properties. This representation: *i*) makes relational patterns directly queryable via SPARQL; *ii*) enables seamless integration with external RDF datasets; *iii*) preserves the portfolio nature of students choices, rather than reducing them to a single indicator.

Interpretable Rule Learning and Neuro-Symbolic Approaches. This thesis will develop hybrid neuro-symbolic methods combining *symbolic rule mining* (concise and interpretable rules), *KG embeddings* [3]. A key challenge is to generate contrastive and counterfactual explanations from heterogeneous models, placing the thesis at the frontier of XAI research [1, 6].

Scalability and Counterfactual Simulation. A further contribution is the development of scalable transformation pipelines for the complete PSupDB and a counterfactual simulation engine within the KG framework, enabling queries such as: "What if this student had been placed in a class with a different social composition?".

2.2. Economics Axis: Causal Identification and Policy Evaluation

Econometric Strategy. Estimating peer effects from observational data is methodologically challenging due to endogeneity: students are not randomly assigned to schools or classes. The thesis will apply multilevel modelling at individual, classroom, and school levels to disentangle peer composition effects from individual characteristics.

Heterogeneity Analysis. Building on subgroup patterns identified by the KG rule-mining component, the thesis will estimate heterogeneous treatment effects across student types.

Policy Simulations. Using the counterfactual infrastructure developed in the CS axis, the economics component will simulate the effects of alternative composition policies.

2.3. Expected Outcomes and Impact

- **Scientific:** publications at the interface of AI/economics across CS venues and economics journals.
- **Empirical:** the first large-scale analysis of peer effects on application portfolios in a national higher education system, covering all tracks including vocational and technological baccalaureates.
- **Methodological:** a reusable, open-source KG framework for PSupDBv alongside validated neuro-symbolic models made publicly available via CASD for authorised researchers.
- **Policy:** concrete, evidence-based estimates of how changes in school and class composition could reduce social and gender inequalities in higher education access.

3. PhD supervision

Christophe Marsala is a full professor in computer science at Sorbonne Université, and a member of the LIP6. His research interests are explainable AI, fuzzy machine learning, and reasoning models. In particular, he develops new approaches for interpretable and explainable decision support models.

Maria Rifqi is a full professor of computer science at the University Paris Panthéon-Assas. As a member of an interdisciplinary lab (the LEMMA brings together researchers in economics, mathematics, and computer science), she has a strong experience in close collaboration with economists. She is also an associated researcher of LIP6. She was coordinator of two scientific projects involving economists: LORIET (2020-2024)¹ and GENREFP (gender career inequalities in the Public Service). She is an expert in ML and Parcoursup data.

Sandra Cavaco is an associate professor of economics (with HDR), expert in applied microeconomics and education economics, member of the LORIET project (2020-2024).

References

- [1] I. Baaaj, Z. Bouraoui, A. Cornuéjols, T. Denœux, S. Destercke, D. Dubois, M.-J. Lesot, J. Marques-Silva, J. Mengin, H. Prade, S. Schockaert, M. Serrurier, O. Strauss, and C. Vrain. Synergies between machine learning and reasoning - an introduction by the Kay R. Amel group. *International Journal of Approximate Reasoning*, 171:109206, 2024.
- [2] R. Bifulco, J. Fletcher, and S. Ross. The effect of classmate characteristics on post-secondary outcomes. *American Economic Journal: Economic Policy*, 3:25–53, 2011.
- [3] P. Hitzler, F. Bianchi, M. Ebrahimi, and M. Sarker. Neural-symbolic integration and the semantic web. *Semantic Web*, 11:3–11, 2020.
- [4] J. Jonsson and C. Mood. Choice by contrast in swedish schools. *Social Forces*, 87(2):741–765, 2008.
- [5] G. Manzo. *La spirale des inégalités*. Presses de l'Université Paris-Sorbonne, 2009.
- [6] T. Pontoizeau, R. Caillière, M.-J. Lesot, C. Marsala, N. Museux, and J.-N. Vittaut. Assessing the Interpretability of Fuzzy Rule Bases. In *IEEE Int. Conf. on Fuzzy Systems*, pages 1–6, Reims, France, 2025.
- [7] B. Sacerdote. Peer effects in education: How might they work, how big are they? In *Handbook of the Economics of Education*, volume 3, pages 249–277. 2011.

¹<https://www.casd.eu/project/loriet-la-plateforme-parcoursup-au-service-de-la-loi-orientation-et-reussite-des-etudiants/>