International Inequalities Institute

WORKING PAPER 150

The network of injustice: A novel approach to inequality of opportunity

Francesco Colcerasa
Italian Ministry of Economics and Finance

Lorenzo Giammei
Italian National Research Council

Francesca Subioli
Department of Law, Roma Tre University

FEBRUARY 2025

Francesco Colcerasa
Italian Ministry of Economics and Finance

Lorenzo Giammei
Italian National Research Council

Francesca Subioli
Department of Law, Roma Tre University

In addition to our working papers series all these publications are available to download free from our website: www.lse.ac.uk/III

International Inequalities Institute
The London School of Economics
and Political Science, Houghton Street,
London WC2A 2AE

E Inequalities.institute@lse.ac.uk
W www.lse.ac.uk/III
X @LSEInequalities

International Inequalities Institute

The network of injustice: A novel approach to inequality of opportunity

Francesco Colcerasa[*] Lorenzo Giammei[†] Francesca Subioli[‡]

Abstract

Restoring the theoretical foundation of John Roemer's conceptualization of inequality of opportunity (IOp), we introduce an innovative empirical approach to measure unfair inequalities through Bayesian networks. This methodology enhances our understanding of income inequality through structural learning algorithms, generating an IOp index and, most importantly, shedding light on the underlying income formation process. We demonstrate how this proposal relates to established measurement methods through simulated data, and provide an application to five European countries to illustrate the potential of Bayesian networks in the context of measuring inequality of opportunity.

JEL Codes: A13, C43, D63, I24
Keywords: Inequality of opportunity, Bayesian networks, EU-SILC

We are grateful to Paolo Brunori, Francisco Ferreira, Pedro Salas-Rojo, and other members of the International Inequalities Institute for their precious comments on this work. We would also like to thank the Global Equality of Opportunity Estimates (GEOM) team, and in particular Paolo Brunori, Pedro Salas-Rojo, and Pedro Torres-Lopez, for the codes made available for inequality of opportunity estimates using machine learning methods. All mistakes in using the codes are our responsibility. Finally, we thank Vito Peragine, Domenico Moramarco and other participants of the 2024 Workshop on Wealth Inequality, Intergenerational Mobility, and Equality of Opportunity in Vienna for their comments. The views expressed in the article are those of the authors and do not necessarily reflect those of the Italian Ministry of Economics and Finance and the Italian National Research Council.

[*]Italian Ministry of Economics and Finance. Contact: francesco.colcerasa@mef.gov.it
[†]Italian National Research Council. Contact: lorenzo.giammei@cnr.it
[‡]Department of Law, Roma Tre University. Contact: francesca.subioli@uniroma3.it

1 Introduction

The distributive issue has long fascinated scholars, and its appeal still makes it one of the most disputed topics in economics. The main controversy is whether and to what extent inequality in a relevant outcome, typically income, but also health, education or wealth, should be tolerated. The economic literature has attempted to address this issue through two conceptually opposed approaches. The first one judges the acceptability of social inequalities on the basis of their consequences on other valuable social outcomes such as economic growth or well-being, i.e. for instrumental reasons (examples are Alesina and Perotti, 1996; Galor and Zeira, 1993; Alesina et al., 2004; Berg et al., 2018; Ferrer-I-Carbonell and Ramos, 2014). On the contrary, a second approach disregards the possible consequences while asking what and how much inequality is *in itself* acceptable according to a criterion of social justice. The distinction is crucial when it comes to justifying redistributive policies: in the former case, they aim to mitigate the bad consequences of neutral social and economic phenomena. In the latter, they aim to apply a criterion of distributive justice that the society cannot otherwise meet.

Universal education can be motivated by reasons of efficiency through the widespread acquisition of human capital and neighbouring effects (Friedman, 1982). Universal basic income can find its social justification from the need to avoid the consequences of poverty on children's development. Taxing wealth at the top can be viewed as necessary to avoid the extreme consequences, up to revolt, resulting from the social unrest of the part of society left behind. Even the wealthiest may claim for less inequality when they fear its consequences such as decreased national productivity or increased crime. On the other side, public policies may find their motivational roots in a requirement for equality of opportunity (Scanlon, 2018). People may be more supportive of policies to reduce inequality, regardless of their income, if they believe that existing inequality is the result of circumstances beyond individual control, i.e. if they perceive it as, at least partially, unfair. Indeed, preferences for inequality-reducing policies can be affected by these two opposing views on how to assess inequality (Fehr et al., 2024; OECD, 2021).

The distinction between morally acceptable and morally unacceptable inequality originated from philosophical egalitarian thought (Rawls, 1958, 1971; Dworkin, 1981a,b; Arneson, 1989; Cohen, 1989), challenging the welfare egalitarian criterion according to which social welfare can depend only on the utility levels of individuals, i.e. on their outcome. A new approach to egalitarianism proposed to insert *personal responsibility* into the discussion of what inequalities are fair, shifting the focus from the outcomes to the sources. The disparity in outcomes among individuals can be considered ethically fair as long as people can be held accountable for it. In other words, a fair distribution of outcomes should reflect individuals' deliberate and conscious choices. The economic discipline has embraced this shift in perspective starting from Sen (1980) and Roemer (1993, 1998), and a rich theoretical and empirical literature aimed at defining and measuring equality of opportunity has flourished (for a comprehensive review of the approaches, see Roemer and Trannoy, 2015, 2016; Ramos and Van de Gaer, 2016; Ferreira and Peragine, 2016). Under this responsibility-sensitive egalitarianism, the differences stemming from factors beyond individuals' control, called *circumstances* in John Roemer's works, are considered morally unacceptable. In contrast, those due to *effort* are fair and should not be addressed by policy. Of course,

there is no clear-cut separation between circumstances and effort in reality, and this represents the main challenge of equality of opportunity measurement.

The main rationale of Roemer's proposal consists of comparing the average outcomes of different groups of people who share the same circumstances, called *types*. Inequality of opportunity can then be identified as the share of inequality in outcomes attributable to those circumstances. Symmetrically, the share of inequality due to individuals' responsibilities – choices – is inequality of effort. However, empirical practice has revealed that drawing this distinction is extremely challenging. Many empirical strategies have been developed to operationalize the philosophical and theoretical formulation of Roemer's concept of inequality of opportunity (among others, see Ferreira and Gignoux, 2011; Checchi and Peragine, 2010; Ramos and Van de Gaer, 2016; Andreoli et al., 2021; Brunori and Neidhöfer, 2021). Any empirical application relies on a crucial theoretical distinction between the *ex-ante* and the *ex-post* approaches. While the former approach compares opportunity sets across individuals using circumstances only, the second approach obtains inequality of opportunity from a rank comparison across circumstance-groups, where rankings approximate effort profusion.[1] Their common rationale is to simulate an income distribution generated only by circumstances, so that the emerging disparities can be unambiguously traced back to unfair differences.

The literature has made many advancements in addressing crucial issues such as circumstances selection and the choice of the best prediction model to simulate the counterfactual income distribution, also making use of modern machine learning techniques (Carranza, 2023; Brunori and Neidhöfer, 2021; Brunori et al., 2024). Among these advancements, particularly relevant are the recent contributions of Carranza (2023) and Brunori et al. (2024), where the issues of downward and upward biases in IOp estimates are remarked, and by Brunori and Neidhöfer (2021) – who deal with these specific issues by implementing machine learning algorithms to perform an optimal model specification. We continue in the same vein as this literature with the additional goal of recovering the original emphasis on measuring inequality of opportunity to guide distributive policies. To this end, we need to return to the roots of the theoretical framework of inequality of opportunity, which focused primarily on the mechanisms underlying unfair inequalities and only secondarily on their actual measurement. This emerges clearly in the contribution by Fleurbaey and Schokkaert (2009) on the importance of structural modelling in the context of equal opportunity analysis, as paraphrased by Roemer and Trannoy (2015):

> «[...] any equality of opportunity empirical analysis must be preceded by an estimation phase to discover the best structural model leading to the results. Only in the second step should we be interested in measuring inequality of opportunity as such.» (Roemer and Trannoy, 2015, p, 273-274)

While most of the advancements in the inequality of opportunity literature focused on the second step of the challenge – better measuring unfair differences by refining the way Roemerian types are defined, we try to recover the first one and go back to structural modelling. In the context of income inequalities, our methodological proposal emphasizes the role of channels transmit-

[1] For more details, see Fleurbaey and Peragine (2013)

ting unfair inequalities throughout the income formation process. Our main goal is to inform policymakers about the engines of inequality of opportunity so that they can more appropriately target distributive policies, while preserving the technical improvements of the most recent literature. To this aim, we propose an innovative methodology to extract inequality of opportunity, able to compute an index of dispersion and detect the structure of the income allocation process. The method draws from both theory-based and data-driven approaches. Specifically, it relies on a probabilistic graphical model called Bayesian network (BN). On the one hand, BNs allow to extract and graphically represent the relationships among circumstances and the outcome from data through machine learning. On the other hand, by means of conditional probability distibutions, they can generate a standard predicted counterfactual distribution of incomes and, thus, an inequality indicator. The proposed perspective enhances the understanding of the data generating process underlying the market remuneration of circumstances, enabling an in-depth analysis of how unfairness shapes the income distribution.

The remainder of the papers is as follows. Section 2 proceeds with a comparison of our proposal with the established techniques for IOp estimation. Section 3 provides the fundamentals of Bayesian networks and a detailed description of the methodological proposal, including a simulation exercise, compared with the most widely used techniques. Section 4 reports an application to five European countries, namely France, Italy, Sweden, Germany, and Poland. Section 5 remarks on the contribution of the article and summarizes the empirical findings, emphasizing its relevance for policy action.

2 Comparing approaches

The theoretical formulation of Roemer (1998) laid the foundation for a new perspective on inequality by linking the discussion of the ethical acceptability of inequality in outcomes to the kind of sources that generate it. As briefly summarised in the introduction, many methods have been proposed to operationalise Roemer's formulation of inequality of opportunity (IOp). In what follows, we provide a brief explanation of the methods most relevant to our goals. In the very first phases of Roemer's work, a Structural Equation Model (SEM) seemed to be the best tool to root the measurement of equality of opportunity on the structure of the income formation process. SEMs allowed to consider the interactions between circumstances, as well as the mediation of effort variables for the effect of circumstances on the outcome, providing a theory-founded structure for the emergence of inequalities. However, some major obstacles prevented researchers from applying this method – in particular, the impossibility of accurately distinguishing effort and circumstance variables – while preferring a reduced-form approach with only circumstances on the right-hand side of a single income equation. With this approach, the coefficients estimated in a linear regression for each circumstance also capture the mediated effect of that circumstance through effort variables.

This reduced-form model – a linear regression of income on circumstances – has been established in empirical practice from the contribution of Bourguignon et al. (2007) on. The starting point for measuring IOp this way consists of simulating a counterfactual distribution of the outcome as if only circumstances were relevant in shaping it. This can be done either parametrically or

non-parametrically (Ferreira and Gignoux, 2011). In the first case, a log-income regression is performed and a parametric linear structure is imposed to the relationship between income and circumstances. Then, a counterfactual distribution is obtained by predicting individual incomes through the estimated coefficients, and the value of an inequality index applied to such predicted outcomes is the measure of inequality of opportunity. By construction, inequality of effort can be obtained as the residual inequality.[2] On the other hand, non-parametric methods are based on groups – the "types" – that are created starting from all possible combinations of circumstances' categories. The average (or median) income of each group is taken as the representative outcome of that particular combination of circumstances. Inequality of opportunity emerges as the variability of such representative incomes, namely as the disparity among representative individuals for each possible opportunity set.

Besides this general framework, many specific improvements have been implemented over the years to address two main issues in measuring IOp: the downward bias due to unobservable circumstances (Ferreira and Gignoux, 2011), and the upward bias due to model overfitting when adding as many circumstances as possible as well as their interactions (Carranza, 2023; Brunori et al., 2024). Some authors, as Brunori and Neidhöfer (2021) and Brunori et al. (2024), remarked on the importance of optimal circumstances selection when creating groups, proposing the adoption of machine learning techniques like *conditional trees*, *transformation trees*, and *random forests* . Table 1 summarizes the similarities and differences between the mentioned methods. While the SEM aims at modelling income with a structure driven by theory and fixed parametrization, the reduced-form regression and the random forest share the predictive vocation. While the relations in the regression are given by theory and the parametrization is fixed, in random forests circumstances' types emerge through iterative splits in "trees", and the parametrization is flexible.

Table 1: Models for IOp measurement

Model	**Main scope**	**Identification of Types**	**Parametrization**
SEM	Modelling	Given by theory	Fixed
Reduced-form	Prediction	Given by theory	Fixed
Random Forest	Prediction	Derived through iterative splits in trees	Flexible
Bayesian Networks	Modelling	Learned through structural learning algorithm	Flexible

Note: The table classifies four possible models for IOp measurement based on their main scope, the nature of the relations they are based on, and the kind of parametrization they rely on. *Source*: Authors's elaborations.

In this context, Bayesian networks represent a viable and good compromise: its main scope is modelling the income generation process, as for SEMs, but the empirical procedure is data-driven and the parametrization is flexible, as in trees and random forests. The main difference with the latter lies in the method's objective. In fact, machine learning techniques – and, specifically, random forests and trees – focus on prediction. Instead, a Bayesian network maps relations among

[2]Notice that the decomposition is exact only adopting an additively decomposable dispersion measure, like those belonging to the Generalized Entropy family (Shorrocks, 1980).

variables employing a graph and a joint probability distribution associated to the graph, providing detailed information on how connected variables affect each other. Remarkably, if the relational structure underlying a set of variables is unknown (as it is in most cases), it can be retrieved from the data and encoded in a graph through machine learning procedures. Importantly, a Bayesian network can also be represented through a SEM (Pearl, 2000), as it shares with it the same objective of representing a system of structural relationships among variables. The research process is reversed but the two-step procedure suggested by Fleurbaey and Schokkaert (2009) is followed: machine learning algorithms learn the structure through a non-parametric data-driven process – maintaining the benefits of ML techniques –, and returns it as a SEM, instead of the researcher establishing the structure and then estimating the parameters. Then, the emerging relations can be used to classically estimate inequality of opportunity. In the next section the methodological details of the model will be described and the BN-based IOp estimands will be compared to those obtained through alternative approaches.

3 Our proposal

In this work, we propose the adoption of Bayesian Networks (Pearl, 1995) to estimate IOp and obtain the map of the relationships between the variables that generate it. The analysis begins with extracting information from a dataset and translating it into a graphical form. This is achieved through structural learning algorithms, namely machine learning procedures that investigate the multivariate relational structure between variables and encode it into a graph. Any available previous knowledge concerning variable interaction can be introduced in the learning phase in the form of constraints. Marginal and conditional probability distributions of the variables are then estimated employing the same dataset according to the structure of the graph. The mentioned steps produce a Bayesian network that can be interpreted as a model of the outcome generating process. The resulting structure can in fact reveal how variables affect one another and contribute to shaping a certain income level. The emerging relationship structure is a crucial contribution of our work, since it allows the researcher to go beyond IOp measurement by shedding light on its generating process. The resulting BN is then exploited to generate a predicted distribution of the outcome that can be used to compute IOp, following what is performed in the standard IOp measurement literature.

The following subsections focus on the technical details of the proposed methodology. In particular, subsection 3.1 provides an accurate description of the methodological fundamentals underpinning Bayesian networks. Subsection 3.2 proposes a simulation exercise, where we compare IOp values estimated through the proposed methodology to those obtained from existing approaches, at increasing sample sizes.

3.1 The statistical model: Bayesian networks

Widely employed in epidemiology, computer science, and some social sciences, Bayesian Networks (Pearl, 1995) are a powerful multivariate statistical tool that is still uncommon in eco-

nomics.[3] Here we introduce the foundational elements of a Bayesian network model and outline the machine learning procedures that allow learning a BN from data.

Definitions A Bayesian network (BN) is a statistical model that consists of a *directed acyclic graph* (DAG) and a *joint probability distribution* over its nodes. A *graph* $G = (V, E)$ is a collection of vertices or nodes V and edges E, and each vertex is associated with a random variable X_i. The edges represent relationships between random variables, and an edge that originates from a node X_i and goes to another vertex X_j is referred to as a *directed edge*. A graph that only contains directed edges is called a *directed graph*. When two nodes are connected by an edge, they are *adjacent* nodes.

A sequence of connected nodes that starts at node X_i and ends at node X_j, regardless of the directions of the edges, is called a *path*. In a *directed path*, all the edges are oriented in the same direction along the path. A directed path from X_i to X_j, where $X_i = X_j$, is referred to as a directed *cycle*. A directed graph is *acyclic* if it does not contain any directed cycle, meaning that there are no paths starting from a node and ending at the same node following the directions of the arrows. BNs literature often adopts kinship terminology to describe relationships among nodes, according to the graph's structure. Specifically, when a directed edge goes from node X_i to node X_j, X_i is denoted as the *parent* of X_j, and conversely, X_j is called the *child* of X_i. Similarly, if a directed path starting from node X_i to node X_j exists, then X_i is said to be an *ancestor* of X_j and of every other node located between X_i and X_j in the directed path.

The joint probability distribution associated to a Bayesian network G with node set $\mathbf{X} = \{X_i, ..., X_n\}$ can always be expressed as:

$$P(x_1, \ldots, x_n) = \prod_i P(x_i|pa_i), \tag{1}$$

where pa_i indicates the *parent set* of variable X_i in G. The product in Equation (1) derives from combining the chain rule of probability calculus with the assumption that the conditional probability of a variable X_i is not sensitive to all its ancestors but only to the variables belonging to its parent set in the DAG (Pearl, 2000). In other words, it is assumed that a variable X_i is independent from all its other ancestors, if we know the value of its parents. A joint probability distribution must admit the factorization of Equation (1) in order to be associated to a Bayesian network. If a probability function P can be factorized as in Equation (1) relative to a DAG G, then it is said that G satisfies the *Markov property*, and P is Markov with respect to G. As a consequence, given a variable X_i belonging to the node set of a Bayesian network, its parent set is sufficient for determining the probability distribution of X_i.[4]

Learnings process When a DAG's structure is unknown, data-driven *structural learning algorithms* can be employed to retrieve the structure of the DAG (Kitson et al., 2023). These algorithms use a dataset as input and, under specific conditions, provide a DAG or a set of DAGs as output. Structural learning algorithms fall into three main families: *constraint-based*, *score-based*,

[3] For a review of how the approach could contribute to the econometric literature, see Hünermund and Bareinboim (2023).

[4] It follows that, once the parent set of a node is established, it is possible to obtain a predicted value of the variable associated to that node employing its conditional distribution given its parents.

and *hybrid*. Constraint-based algorithms reveal the graph's structure by examining conditional independence relationships in the data. Typically, they start with a fully connected graph and iteratively check for marginal or conditional independence between variables. If independence is found, the corresponding edge is removed. On the other hand, score-based algorithms rely on a scoring function to assess how well a DAG reflects the relations among the variables of the dataset. They start by computing the score of an initial graph structure, and then modify the graph by introducing, deleting, or reversing edges. The final output is the configuration with the highest score. Finally, hybrid algorithms aim to combine the strengths of both score-based and constraint-based approaches. This integration enhances the robustness of structural learning, leveraging both methodologies.[5]

If some prior knowledge concerning the subject matter is available, structural learning algorithms can also account for this information during the learning procedure. Incorporating domain-specific knowledge reduces computational time and makes the graph more credible, increasing the interpretability of the resulting network. In particular, prior knowledge can be incorporated in the model in the form of *forced* or *forbidden* arcs, i.e. of directed relations which must be present in the DAG, or others which are not allowed during the learning process. In economic applications such as ours, this feature allows all information from economic theory to be incorporated into the model, as well as additional modelling features – like the temporal sequencing of variables – that make the output more credible in the specific context. Once the structure of the graph has been learned, the parameters defining the conditional probability distribution associated to the BN are estimated through the EM algorithm (Moon, 1996).

3.2 A simulation exercise

To validate our proposal for measuring inequality of opportunity, we believe it is necessary to verify that, at least for simple data generation processes, the new methodology generates IOp values consistent with those emerging from previously established approaches. Therefore, in this section, we provide the results of a simulation exercise aimed at comparing the performance of BNs with respect to some existing approaches. The simulation process starts by extracting samples of increasing size (from 100 to 20,000 observations, with a step of 400) according to a data generating process described in panel (b) of Figure 1 and a set of conditional probability tables available in Appendix tables A.1-A.4.[6] According to this data generating process, the father's education affects his occupation, which directly affects both the financial distress of the household when the respondent was 14 years old and the labour earnings of the child when adult. The financial distress of the household is assumed to have no influence – either direct or mediated – on income. The simulated dataset will therefore reflect all the distributional features encoded in this simple model.

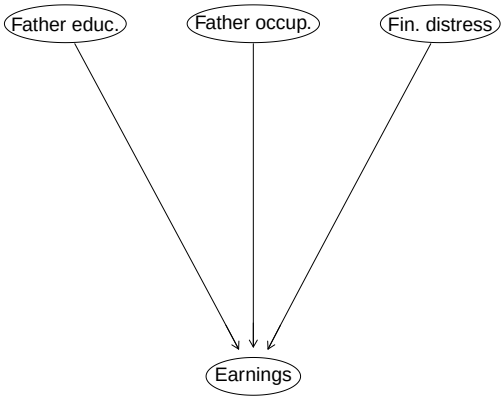
We proceed by measuring IOp on the simulated data employing five different methods. The

[5]The most employed constraint-based algorithm is the *PC algorithm* (Glymour et al., 1991). Algorithms in the score-based category include the *greedy search*, *simulated annealing*, and *genetic algorithms* (Russell and Norvig, 2016). Examples of hybrid algorithms include *Max-Min Hill Climbing* (Tsamardinos et al., 2006) and H^2PC (Gasse et al., 2014).

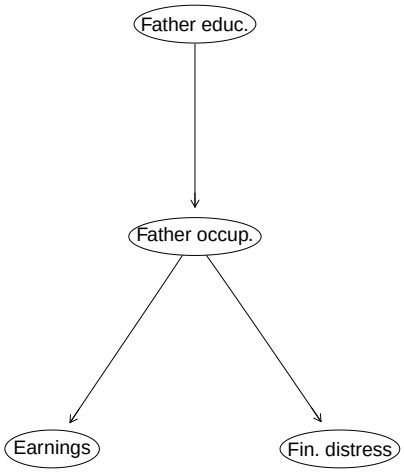
[6]We use for the simulation the marginal and the conditional distributions estimated through EU-SILC data for Italy in 2019 with the appropriate sample weights.

Figure 1: Examples of simple data generating processes

(a) DGP underlying a linear regression



(b) More complex DGP



Note: The figure plots two possible data generating processes (DAG) for labour earnings. The network in panel (a) mimics the data-generating process assumed by a simple linear regression model with no interaction.

first one follows the traditional approach implying a reduced-form linear regression of labour earnings on each circumstance with no interaction (panel a of Figure 1). The second method employs a Bayesian network that mimics the reduced-form linear regression by imposing again the structure of panel (a) of Figure 1. This allows to study the behaviour of a BN when it is built to be as close as possible to a linear regression model. The third, most crucial, method employs a Bayesian network obtained by applying a structural learning algorithm to the simulated data: this corresponds to the procedure we propose in this work. To learn the network, we employ a score-based algorithm called "Tabu Search" (Russell and Norvig, 2016) with a BIC score.[7] The counterfactual distribution of the income is then predicted according to the conditional distribution of the income given its parents, according to the Markov property embedded in Equation (1). Finally, the fourth and fifth methods follow the most recent IOp empirical literaure by employing *Conditional Inference Regression Trees* and *Conditional Inference Random Forests* as developed by Hothorn et al. (2006) and applied in Brunori et al. (2023).[8] A Conditional Inference Regression Tree is a supervised machine learning algorithm aiming at partitioning the regressors' space to predict the variability of a dependent variable. The population is partitioned by recursive binary splitting of the sample into an exhaustive and mutually exclusive set of subgroups (the Romearian types in this context). Given the high variance and strong sample dependence of conditional trees, a random forest approach can be used to generate robust estimates. A Conditional Inference Random Forest draws different subsamples of the original data, and computes a tree on each one.[9] IOp is measured in a second step for all the five estimation approaches by computing

[7]While other choices are possible, Tabu Search is faster and more accurate than most algorithms for both small and large sample sizes (Scutari et al., 2019). Note that, even if the BN that mimics regression has an imposed structure while the other one has a learned structure, in both cases the conditional distribution parameters are estimated on the simulated dataset employing the EM algorithm.

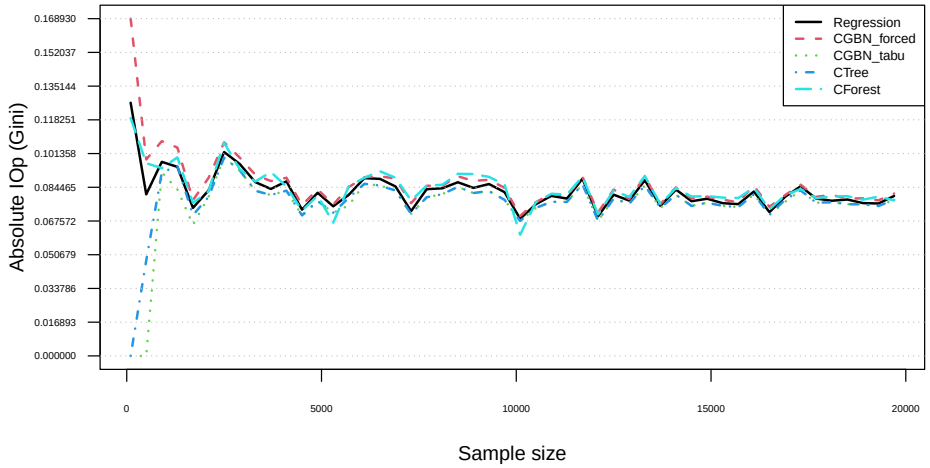
[8]We are grateful to the Global Equality of Opportunity Estimates (GEOM) team for material, code, and support for implementing these methods.

[9]Following Hothorn et al. (2006), we set alpha to 1; the share of each subsample drawn in every iteration is the default 0.632, and the minimum number of observations allowed in each terminal node is 0.1% of the sample size (or

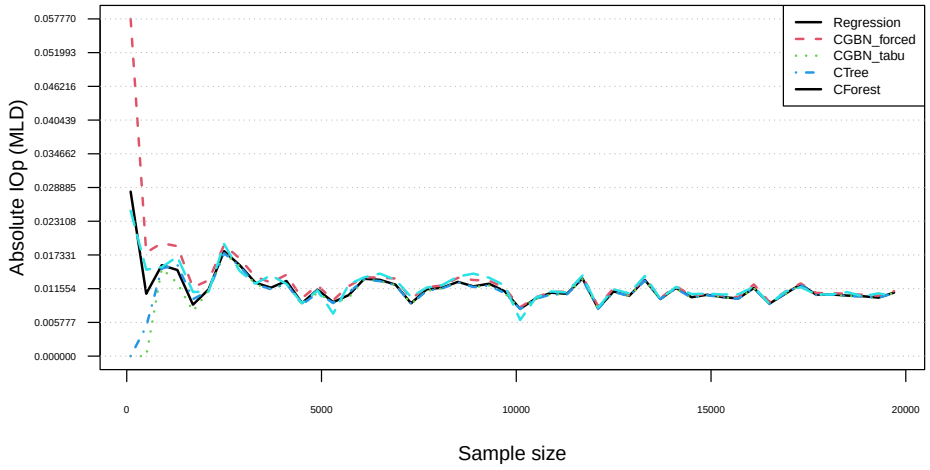
the Gini index and the Mean Logarithmic Deviation on the predicted income distribution. The results are compared in Figure 2.

Figure 2: Convergence of inequality of opportunity indices

(a) Gini Index



(b) Mean Logarithmic Deviation



Note: The figure plots the inequality of opportunity indices (Gini index in panel a, mean logarithmic deviation in panel b) computed on the predicted earnings obtained through a reduced-form linear regression with no interactions ("Regression"), a Bayesian network that assumes the same relational structure of that regression ("CGBN_forced"), a Bayesian network with a structure learned from data through a structural learning algorithm ("CGBN_tabu"), a Conditional Inference Regression Trees ("CTree"), and a Conditional Inference Random Forest ("CForest"). The underlying data generating process is the one in panel (b) of Figure 1, and the probability distributions are those for Italy as reported in Tables A.1-A.4 in the Appendix. Random samples of increasing size are drawn with a step of 400 from 100 to 20,000 observations. *Source*: Authors's elaborations on simulated data based on IT-SILC data for 2019.

The two plots highlight how the results of the four methods converge as sample size increases for both the Gini index (panel a) and the MLD (panel b). These patterns confirm how, at higher sample sizes, the methods produce almost the same IOp estimates. This is consistent with the idea that the employed statistical techniques share the same rationale of finding the best way of

10, if the sample size was smaller than 1000). All remaining tuning parameters are set to the default values in the "cforest" R function in "partykit" (Hothorn et al., 2006).

predicting the outcome given a set of covariates, under the assumptions implied by each model. Interestingly, for small samples the learned Bayesian network and the Conditional Tree measure zero to low levels of inequality: both methods cannot capture heterogeneity in the data when the sample is too small, and predict only one income value for all the observations.[10] With increasing sample size, they converge to the estimates of the other methods and are very close to each other. This simple exercise is useful to clarify the main point of this work: the reason we are trying to return to modelling in measuring IOp is related to a policy objective. We believe that providing a number, or a ranking, or a time trend, is relevant but not sufficient. Policy needs to know *how* that number is generated in society, that is, the process behind the formation of inequality of opportunity. Without any claim on causality, which is unattainable in such complex contexts and with the current data availability, we still think structural learning may be the key.

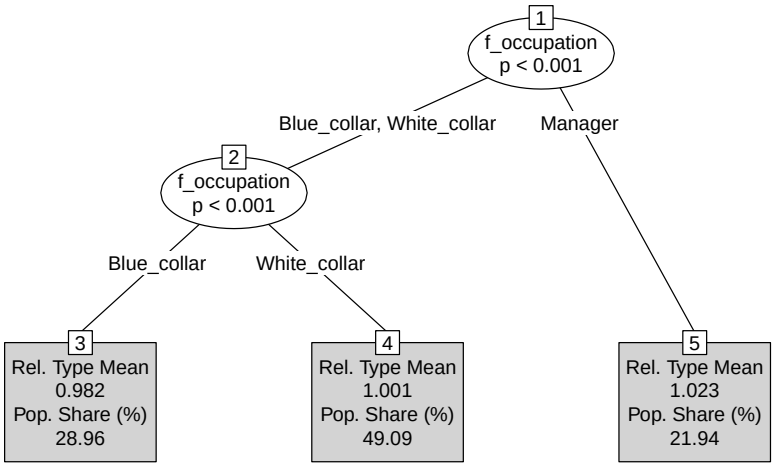
To illustrate the proposal's potential in this direction, we show below in Figure 3 the tree generated in the simulation for the largest sample numerosity (n = 20,000). We see that the machine learning algorithm operates two sequential splits: a first one for paternal occupation, separating managers from the other two occupational categories, and a second one separating individuals with blue-collar and white-collar fathers. The result is a categorisation in three types with different paternal occupations. The tree is consistent with the generative process of the data in panel (b) of Figure 1, and with the conditional earnings by paternal occupation in Table A.4 in the Appendix. The comparison of this conditional tree with the network in Figure 1 is particularly interesting for our purposes: it allows us to highlight the complementarity of the two approaches. While providing perfectly comparable estimates of IOp (see Figure 2), the two graphs contain different information. The conditional tree allows us to visualise the types (in Roemer's sense), and the result of the independence tests between the groups. On the other side, the network in panel (b) of Figure 1 reveals the relationships between the variables, and thus the structure of the influence of circumstances on income. The difference is crucial from a policy point of view. Indeed, the information in Figure 3 makes it possible to identify the categories of individuals or households that have a relatively worse outcome due to circumstances beyond their control, and thus target assistance policies to those groups. Besides, the information from the network allows to understand which mechanisms have generated the disadvantage, to be able to intervene "at the roots" of inequality of opportunity.

4 An application to five European countries

We implement an empirical application of our methodological proposal over five European countries (Poland, Germany, Italy, France, and Sweden) reflecting different configurations of welfare systems – *conservative* for Germany and France, *familistic* for Italy, and *socio-democratic* for Sweden. Such a distinction partially follows the classification of welfare systems in Esping-Andersen (1990) distinguishing *liberal*, *conservative*, and *socio-demographic* regimes, but takes into account the peculiarities of the Italian welfare system. Indeed, in addition to the typical corporatism of conservative regimes (like Germany and France), the Italian welfare system relies strongly on the

[10] On this point, we verified that the Conditional Tree generated for n = 100 was made of one single "leaf", i.e. only one Romerian type is generated.

Figure 3: Conditional Inference Tree of simulated data



Note: The figure plots the Conditional Inference Regression Tree generated for a random sample of 20,000 observations drawn form a distribution based on the data generating process in panel (b) of Figure 1 and the probability distributions for Italy as reported in Tables A.1-A.4 in the Appendix. Parametrization: $\alpha = 1$; the share of each subsample drawn in every iteration is the default 0.632; the minimum number of observations allowed in each terminal node is 2,000 of the sample size; all remaining tuning parameters are set to the default values in the "cforest" R function in "partykit" (Hothorn et al., 2006). *Source*: Authors's elaborations on simulated data based on IT-SILC data for 2019 and on the code developed by the GEOM team.

family as a social insurance mechanism. Finally, Poland is itself an interesting case study, being a country of the former Union of Soviet Socialist Republics (USSR) and part of the Eastern European countries. Beyond such considerations, the general aim of the application is to show how a learned Bayesian network looks like in the context of inequality of opportunity analysis, and how it can be used to inform the policy beyond assigning a number to inequality of opportunity.

4.1 Data

The empirical application is entirely based on the 2019 release of the European Union Statistics on Income and Living Conditions (EU-SILC). Specifically, we use the *ad-hoc* module issued in 2019 focusing on the intergenerational transmission of disadvantages. It is part of a sub-set of modules collected periodically (other waves are in 2005 and 2011) that retrieve information on individuals' socio-economic background. The variables included in the module capture the childhood and parental background information of the respondent when he or she was around 14 years old.[11] Specifically, our analysis relies on the following background variables: father's and mother's level of education and type of occupation; the type of household in which the respondent used to live; the number of siblings living at home; the degree of urbanization of the living area; an indicator of perceived financial distress in the household. Moreover, we exploit the cohort of birth, the sex, and the country of birth as demographic circumstances. Table 2 below reports the detailed definition of all the variables included in the model.

[11] Slight differences might arise for different countries. In general, background information is retrieved for an age ranging between 12 and 16 years old.

Table 2: Variable Descriptions

Variable	Type	Definition
Labour earnings	Continuous	Annual earnings measured for 2018 at the individual level, coming from working activities, either as employee or as self-employed, gross of social security contributions and taxes.
Cohort of birth	Categorical	Class of cohort of birth: Baby boomers (1959-1969), X Generation (1970-1979), Millennials (1980-1993).
Sex	Categorical	Indicator variable taking value 1 if the respondent is a man, and 0 if the respondent is a woman.
Country of birth	Categorical	Categorical variable for the location of birth of the respondent: Local, EU, Extra-EU, Other/Unknown.
Education of father	Categorical	Categorical variable capturing the highest level of education of the father when the respondent was around 14 years old: Low (at most primary education), Medium (secondary education), High (tertiary education), Other/Unknown.
Education of mother	Categorical	Categorical variable capturing the highest level of education of the mother when the respondent was around 14 years old: Low (at most primary education), Medium (secondary education), High (tertiary education), Other/Unknown.
Occupation of father	Categorical	Categorical variable capturing the skill level of father's occupation when the respondent was around 14 years old, obtained by combining the activity status and the ISCO-8 classification of the father's occupation in EU-SILC: Unemployed/Inactive, Low-skilled, Medium-skilled, High-skilled, Other/Unknown.
Occupation of mother	Categorical	Categorical variable capturing the skill level of mother's occupation when the respondent was around 14 years old, obtained by combining the activity status and the ISCO-8 classification of the mother's occupation in EU-SILC: Unemployed/Inactive, Low-skilled, Medium-skilled, High-skilled, Other/Unknown.
Household type	Categorical	Categorical variable for the presence of parents in the household when the individual was around 14 years old: Both parents, One parent, No parents, Other/Unknown.
Number of siblings	Categorical	Categorical variable capturing the number of siblings in the household when the individual was around 14 years old: No or one sibling, Two or three siblings, Four or more siblings, Other/Unknown.
Urbanization	Categorical	Categorical variable capturing the level of urbanization of the area the household lived in when the respondent was around 14 years old: City, Town/suburb, Rural area, Other/Unknown.
Financial distress	Categorical	Indicator variable capturing a perceived bad financial situation of the household when the respondent was around 14 years old.

Note: The table reports the type and definition of the variables used in the application. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. *Source*: Authors's elaborations on EU-SILC data 2019.

We restrict the sample to respondents aged between 25 and 59. As the reference notion of income, we employ earnings from labour measured at the individual level in 2018, coming from working activities, either as employee or as self-employed, gross of social insurance contributions and taxes. Unemployed and inactive individuals are included with zero labour earnings, unless they declare to be students, retired, or unable to work due to disability. While the literature usually adopts household equivalised disposable income, we prefer focusing on individual labour income (before redistribution and irrespective of the household structure), being our focus the effect of circumstances on the individual income formation process.[12] To make the sample representative of the country-level population, we appropriately apply the individual sample weights provided with the survey. Table A.5 reports descriptive statistics for the variables included in the model for each country.

4.2 Computing IOp through Bayesian networks

We proceed by learning five Bayesian networks – one for each country – through the "Tabu Search" structural learning algorithm, and then use the obtained networks for measuring inequality of opportunity as explained in the methodological section. The Tabu Search algorithm is implemented employing the extended Bayesian Information Criterion (Foygel and Drton, 2010), which adds a second penalty to BIC to penalize dense networks, i.e., networks with many edges. This score ensures that the three networks have a comparable "density" of relations despite the size of the national samples coming from EU-SILC is different (see Table A.5).

As explained in subsection 3.1, Bayesian networks allow to incorporate previous knowledge on the data generating process into the model through *forced* and *forbidden* arcs. We exploit both economic theory and the temporal sequence of the variables included in the model to reduce the number of edges, relating either circumstances with each other or circumstances with labour earnings. Specifically, we exploit the fact that no variable can influence any preceding one, while allowing all variables to influence any contemporaneous or subsequent variable. Table 3 reports the five groups in which we divide the variables according to the period in which we can reasonably assume that they emerge.

Table 3: Time sequence of variables

Timeline →

Birth	**Parent education**	**Parent occupation**	**Background at age 14**	**Present**
Cohort of birth	Education of father	Occupation of father	Urbanization	Labour earnings
Country of birth	Education of mother	Occupation of mother	Financial distress	
Sex			Household type	
			Number of siblings	

Note: The table reports the temporal line the authors use to order individual circumstances and labour earnings.In learning the Bayesian network on the selected sample (see subsection 4.1), this timeline imposes that no variable can influence any preceding one.

[12]However, we are aware that the approach has limitations due to possible behavioural responses from anticipating the effect of redistribution and planning one's labour supply according to family composition.

The first group is made of the demographic circumstances: the cohort of birth, the sex, and the country of birth. Of course, in a cross-section analysis like ours, the cohort of birth is also perfectly correlated with age. The background variables are divided into three groups: the first one is made of each parent's education, the second one of their occupation, and the last one contains the four variables related to the situation of the respondent's household when he or she was 14-years-old – the degree of urbanization in the area, whether the household suffered a bad financial situation, whether the respondent used to live with both parents or not, and the number of siblings at home. Finally, the last group is in the present (2018 in our application) and consists of labour income only. In general, inside each group in Table 3 any relation is possible; however, the variables in the first group (birth circumstances) cannot influence each other for obvious reasons. Moreover, to simplify the model we prevent the education of one parent from influencing that of the other, assuming that the education decision precedes the household formation. In addition to the time-based constraints, we further add two reasonable *forbidden* and two *forced* relations to the model: each parent's education level cannot affect the occupation of the other parent, while each parent's education level necessarily impacts his or her own occupation.[13]

4.3 Results

Table 4: Inequality of opportunity estimates in 2018

	Country				
Index	Italy	Germany	Poland	France	Sweden
Gini (%)	0.20 (40.8)	0.19 (45.5)	0.16 (35.8)	0.15 (38.2)	0.11 (35.0)
MLD (%)	0.064 (22.4)	0.058 (20.4)	0.043 (18.1)	0.037 (13.0)	0.022 (11.1)

Note: The table reports the estimates of inequality of opportunity (level and percentage of total inequality) as measured by the Gini index and the Mean Logarithmic Deviation through a Bayesian network for France, Italy, Germany, Poland, and Sweden, using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. *Source*: Authors's elaborations on EU-SILC data 2019.

Table 4 reports the estimated levels of IOp for the five countries as measured by the Gini index and the Mean Logarithmic Deviation of labour earnings predicted through the network of circumstances.[14] The estimated levels of IOp are generally in line with previous findings, explaining between 35% and 46% (for Gini) and between 11% and 22% (for MLD) of overall income inequality. However, the most valuable characteristic of using a Bayesian network in this framework is the possibility of looking into the network itself. By their very nature, BNs are readable and the relations that emerge are significant and can be interpreted. Therefore, we look into the structure of the three networks learned for France (Figure 4), Italy (Figure 5), Germany (Figure 6), Poland (Figure 7), and Sweden (Figure 8). The structure of a BN allows to immediately grasp which is the "hierarchy" of circumstances and other variables that influence labour income:

[13]The results are quantitatively identical if we limit the constraints to those that prevent labor income from influencing any circumstance and those that do not allow any variable to influence sex, birth cohort and country of birth.

[14]As explained in Section 3, we remark that labour earnings predictions have been obtained by employing the conditional distribution of labour earnings given its parents, according to the learned DAG structure.

the lower the income variable is in the network, the more complex is its generative process. Also note that some of the arcs appearing in the networks have been introduced as prior knowledge (see subsection 4.2), while others have been prohibited.

Table 5: Summary of parent and ancestor variables of income

	France	**Italy**	**Sweden**	**Germany**	**Poland**
Demographics					
Cohort of birth	Parent	Parent	Parent	Parent	Ancestor
Country of birth	Ancestor	Ancestor	Ancestor	Parent	Ancestor
Sex	Parent	Parent	Parent	Parent	Parent
Background					
Education of Father	Ancestor	Parent	Ancestor		Ancestor
Education of Mother	Ancestor		Ancestor	Parent	Ancestor
Occupation of Father	Parent				
Occupation of Mother	Ancestor		Parent		Parent
Urbanization					
Financial distress					
Household type					
Number of siblings					

Note: The table reports for each country (columns) whether a circumstance (rows) is a "parent" variable of labor income (a direct determinant), an "ancestor" of it (a determinant mediated by other circumstances), or neither. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

Demographic circumstances Table 5 summarizes the direct and indirect relationships with income that emerge for different countries. We begin by identifying in the graphs the "parents" of the income variable in the five countries, that is, the circumstances directly related to income, without mediators.[15] Interestingly, for all countries except Germany, income has only three parent variables, while in Germany it has four. Starting from demographic variables, as expected, sex is in all cases a direct determinant of income.[16] The cohort of birth is also a direct determinant in all countries but for Poland, where its effect is instead mediated by the educational level of the parents and by mother's occupation. The country of birth is a parent variable for income only in Germany, while in Italy its effect is mediated through father's education, in France through both parents' education and occupation, in Poland and Sweden by the educational level of the parents and by mother's occupation.[17]

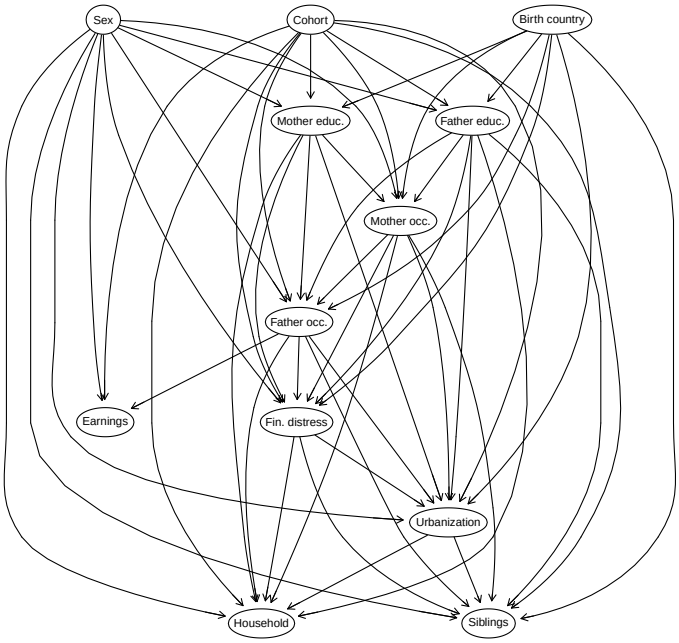
Background circumstances As regards background variables, the education of the father is a direct determinant of income in Italy, and his occupation in France. In Germany and Poland, the

[15]We recall from methodology that, since the learned graph is Markov, it is possible to condense the chains of conditional dependence into the direct parents only: they convey all the information of their "ancestors", i.e. of the variables connected to income through the parent variables.

[16]When interpreting this result, not that we include in the data inactive respondents with zero earnings declaring to devote themselves entirely to care activities. Therefore, the income variable also include information on the activity status.

[17]In our approach, the country of birth is a categorical variable that indicates whether the individual was born in the country, immigrated from another EU country, or immigrated from a non-EU country. The approach is thus different from considering the country of residence as a circumstance itself (Milanovic, 2015), and from the perspective of the European Union as a single country with several regions (Brandolini, 2007).

Figure 4: Bayesian network for France in 2018



Note: The figure plots the Bayesian network for France obtained through the statistical learning process described in section 3 using appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

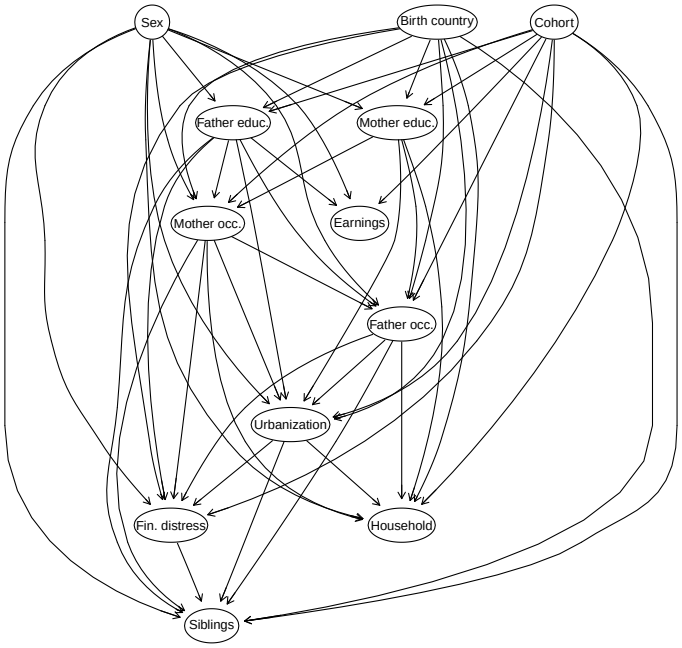
education of the mother plays a direct role, and in Poland and Sweden her occupation. However, in some cases, the other parental education and occupation variables that are not direct determinants of income are nonetheless its "ancestors", that is, determinants mediated by other circumstances. Exceptions are Italy, for which only the father's education matters, and Germany, for which only the mother's education is relevant for income. It is very interesting to note that all other background variables at age 14, except for parental education and occupation, while interrelated, do not play any role in the networks in determining income.

Though some similarities emerge, each country reports a specific income formation process and specific mechanisms of influence for the circumstances. Such comparison should be read in light of IOp values as well. Indeed, even though some countries might report a similar structure of the unfairness' transmission, different levels of IOp might be observed. For example, Italy and Germany have the highest levels of IOp (absolute and relative) in Table 4, but the simplest structure of inequality of opportunity according to the relations in Table 5.

5 Discussion and conclusion

Part of the complex debate on inequality has been characterized by the attempt to justify interpersonal differences according to social justice criteria. The economic literature has made a wide range of contributions on this, starting from the work of John Roemer who provided an

Figure 5: Bayesian network for Italy in 2018



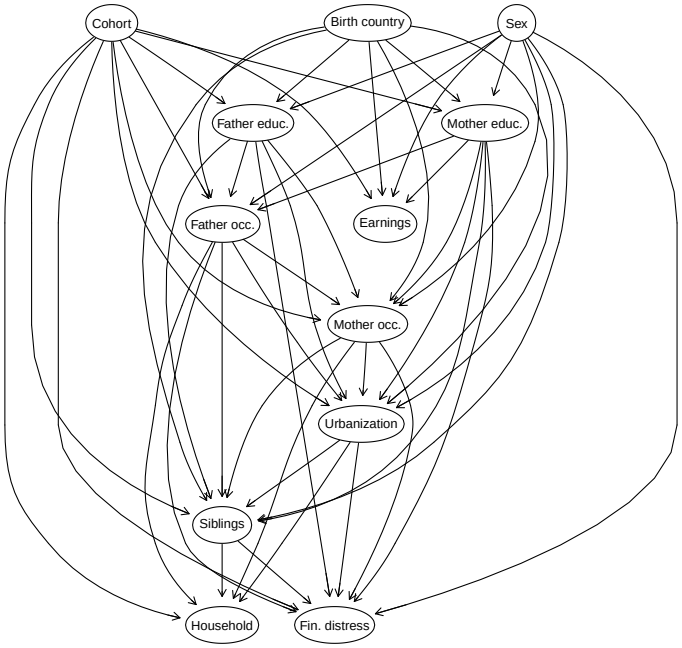
Note: The figure plots the Bayesian network for Italy obtained through the statistical learning process described in section 3 using appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

ethical metric for assessing inequality. According to this view, gaps in individual outcomes (such as income, wealth, etc.) are ethically acceptable to the extent that they emerge from choices for which individuals can be held accountable, while those that emerge from factors beyond personal control should be compensated. In the wake of this distinction, various measurement approaches have been proposed. However, the theoretical spirit of Roemer's contribution has been overshadowed by crucial empirical challenges that have emerged over the years.

In this work, we propose an innovative approach to measure inequality of opportunity that recovers the focus on structural relations underlying the original Roemerian approach, while combining it with the estimation effectiveness of recent empirical approaches. The proposal aims to unravel the complexity of the network of relationships underlying the process linking circumstances to adult income. Moreover, in addition to producing an index of unfairness comparable with previous approaches, it enables to extract policy-oriented evidence. In fact, the adoption of Bayesian networks (Pearl, 1995) allows to dig up the allocation process that generates such unfair inequality. Inspecting the network that connects circumstances with each other and with income, it is possible to identify the channels that activate inequality of opportunity. This innovation makes the methodology informative to the policymakers about which actions can effectively reduce unfair inequalities.

A simulation exercise allowed us to show the complementarity of the proposed method with the most widely used, while an empirical application on five European countries in different welfare

Figure 6: Bayesian network for Germany in 2018

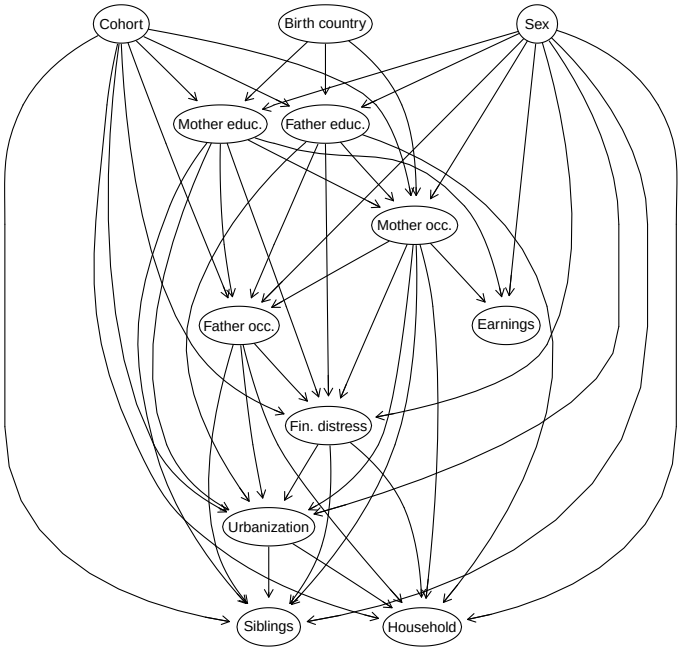


Note: The figure plots the Bayesian network for Germany obtained through the statistical learning process described in section 3 using appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

regimes provided an example of the interpretive potential with real data. In the application, we showed that while the estimates are in line with the literature, the network is also informative about how the background conditions affect downstream outcomes, uncovering similarities and differences across countries. The tool can be very useful for comparing countries but also the same country over time, especially if it has gone through major structural or institutional changes. The graphical representation of the model might represent a key tool in policy-making, as it enables a visualization of the network of injustice, that is, the mechanisms underlying unfair allocations of income that generate inequality of opportunity.

Additional data on circumstances and, possibly, effort would provide more detailed and granular information on the workings of the income formation process underlying inequality of opportunity. The emerging map of relationships among variables would be informative both about the underlying factors relevant to unjust interpersonal gaps and the mechanisms that transmit their effect to final outcomes in the labour market. The network would emerge as an excellent policy tool for addressing inequity in two respects. On the one hand, it would make it possible to understand through what mechanisms (e.g., sector, occupation, labour intensity) background circumstances influence income in the current generation, enabling the policymaker to intervene on those mechanisms. On the other hand, it would ease the understanding of which background characteristics influence income and thus allow to act on the socio-economic endowment of future parents (e.g., by increasing the level of education today).

Figure 7: Bayesian network for Poland in 2018

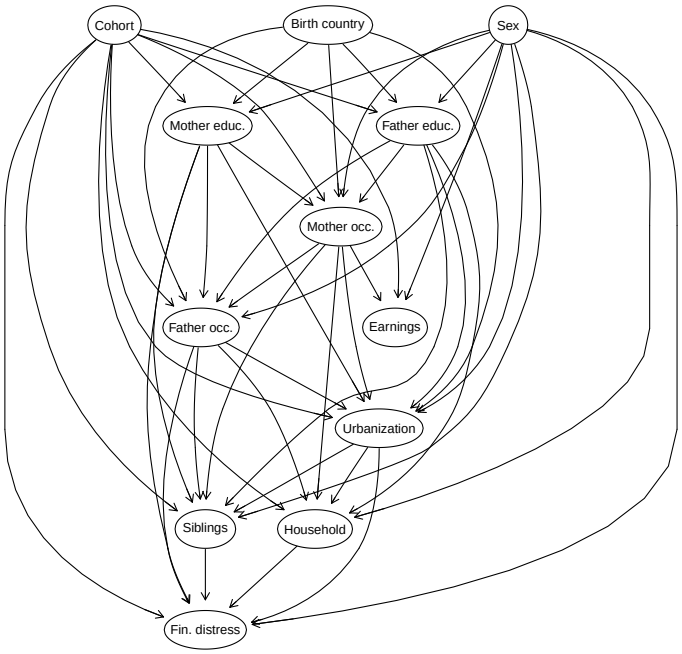


Note: The figure plots the Bayesian network for Germany obtained through the statistical learning process described in section 3 using appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

Two lines of research emerge as a natural continuation of this work. A first one concerns the role of household formation and redistribution in reshaping the map of relations we have shown in the application. A possible method could encompass the adoption of equivalised household income and disposable income after taxation and transfers to learn the graph, and study what happens to the level of inequality of opportunity and to the network when family composition and the tax and transfer system are taken into account. A second possible line of research concerns investigating the effect of circumstances beyond individual control on outcome variables other than income, such as wealth, consumption, but also health. Such an approach may provide a more complete picture of the mechanisms driving broader concepts of inequality of opportunity.

Other possible extensions of the approach concern policy simulation and ethical comparisons. Regarding the first aspect, the proposed methodology is suitable for the evaluation of public interventions. In fact, BNs make it possible to simulate a shock at any node and observe its propagation through all the arcs of the network. This would be a useful tool to test the impact of policy measures aimed at reducing the effects of inequality of opportunity. Secondly, the approach can be used to test the real data generating processes against an existing ethical benchmark, in the spirit of Andreoli et al. (2019). The network learned from the data can in fact be compared with a reference network of equality of opportunity through measures of "distance", which could then be interpreted as a measure of unfairness in society. The definition of the benchmark could be "absolute", reflecting theories of distributional justice, or be "relative", taken from the views

Figure 8: Bayesian network for Sweden in 2018



Note: The figure plots the Bayesian network for Sweden obtained through the statistical learning process described in section 3 using appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. A detailed description of the variables included in the model is available in Table 2 and Table A.5. *Source*: Authors's elaborations on EU-SILC data 2019.

and opinions of a given society. These ideas are left for future research.

References

Alesina, A., R. Di Tella, and R. MacCulloch (2004). Inequality and Happiness: are Europeans and Americans different? *Journal of Public Economics 88*, 2009–2042.

Alesina, A. and R. Perotti (1996). Income distribution, political instability and investment. *European Economic Review 40*, 1203–1228.

Andreoli, F., A. Fusco, I. Kyzyma, and P. Van Kerm (2021). New estimates of inequality of opportunity across European cohorts. *36th IARIW Virtual General Conference*.

Andreoli, F., T. Havnes, and A. Lefranc (2019). Robust inequality of opportunity comparisons: theory and application to early childhood policy evaluation. *Review of Economics and Statistics 101*(2), 355–369.

Arneson, R. J. (1989). Equality and equal opportunity for welfare. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 56*(1), 77–93.

Berg, A., J. D. Ostry, C. G. Tsangarides, and Y. Yakhshilikov (2018). Redistribution, inequality, and growth: new evidence. *Journal of Economic Growth 23*, 259–305.

Bourguignon, F., F. H. Ferreira, and M. Menéndez (2007). Inequality of opportunity in Brazil. *Review of income and Wealth 53*(4), 585–618.

Brandolini, A. (2007). Measurement of income distribution in supranational entities: The case of the European Union. *Bank of Italy Temi di Discussione Working Paper* (623).

Brunori, P., F. H. Ferreira, and P. Salas-Rojo (2024). Inherited Inequality: A General Framework and an Application to South Africa. Technical report, Center for Open Science.

Brunori, P., P. Hufe, and D. Mahler (2023). The roots of inequality: Estimating inequality of opportunity from regression trees and forests. *The Scandinavian Journal of Economics 125*(4), 900–932.

Brunori, P. and G. Neidhöfer (2021). The evolution of inequality of opportunity in Germany: a machine learning approach. *Review of Income and Wealth 67*, 900–927.

Carranza, R. (2023). Upper and Lower Bound Estimates of Inequality of Opportunity: A Cross-National Comparison for Europe. *Review of Income and Wealth 69*(4), 838–860.

Checchi, P. and V. Peragine (2010). Inequality Of Opportunity in Italy. *Journal of Economic Inequality 8*, 429–450.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics 99*(4), 906–944.

Dworkin, R. (1981a). What is equality? part 1: Equality of welfare. *Philosophy Public Affairs 10*(3), 185–246.

Dworkin, R. (1981b). What is equality? part 2: Equality of resources. *Philosophy Public Affairs 10*(4), 283–345.

Esping-Andersen, G. (1990). *The Three Worlds of Welfare Capitalism*. Princeton University Press.

Fehr, E., T. Epper, and J. Senn (2024, 09). Social Preferences and Redistributive Politics. *The Review of Economics and Statistics*, 1–45.

Ferreira, F. and J. Gignoux (2011). The Measurement Of Inequality Of Opportunity: Theory and an Application To Latin America. *Review of Income and Wealth 57*, 622–657.

Ferreira, F. H. G. and V. Peragine (2016, 06). Individual responsibility and equality of opportunity. In M. D. Adler and M. Fleurbaey (Eds.), *The Oxford Handbook of Well-Being and Public Policy*. Oxford University Press.

Ferrer-I-Carbonell, A. and X. Ramos (2014). Inequality and Happiness. *Journal of Economic Surveys 28(5)*, 1016–1027.

Fleurbaey, M. and V. Peragine (2013). Ex ante versus ex post equality of opportunity. *Economica 80*(317), 118–130.

Fleurbaey, M. and E. Schokkaert (2009). Unfair inequalities in health and health care. *Journal of health economics 28*(1), 73–90.

Foygel, R. and M. Drton (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in neural information processing systems 23*.

Friedman, M. (1982). *Capitalism and Freedom*. Chicago: University of Chicago Press.

Galor, O. and J. Zeira (1993). Income distribution and macroeconomics. *The Review of Economic Studies 60*, 35–52.

Gasse, M., A. Aussem, and H. Elghazel (2014). A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. *Expert Systems with Applications 41*(15), 6755–6772.

Glymour, C., P. Spirtes, and R. Scheines (1991). Causal inference. *Erkenntnis 35*(1-3), 151–189.

Hothorn, T., K. Hornik, and A. Zeileis (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics 15*(3), 651–674.

Hünermund, P. and E. Bareinboim (2023). Causal inference and data fusion in econometrics. *The Econometrics Journal*, utad008.

Kitson, N. K., A. C. Constantinou, Z. Guo, Y. Liu, and K. Chobtham (2023). A survey of Bayesian Network structure learning. *Artificial Intelligence Review 56*(8), 8721–8814.

Milanovic, B. (2015, 05). Global inequality of opportunity: How much of our income is determined by where we live? *The Review of Economics and Statistics 97*(2), 452–460.

Moon, T. (1996, November). The expectation-maximization algorithm. *IEEE Signal Processing Magazine 13*(6), 47–60.

OECD (2021). Does inequality matter?: How people perceive economic disparities and social mobility.

Pearl, J. (1995). Causal Diagrams For Empirical Research. *Biometrika 82*, 669–688.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge, Cambridge University Press.

Ramos, X. and D. Van de Gaer (2016). Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys 30*(5), 855–883.

Rawls, J. (1958). Justice as Fairness. *The Philosophical Review 67*(2), 164–194.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Roemer, J. E. (1993). A pragmatic theory of responsibility for the egalitarian planner. *Philosophy Public Affairs 22*(2), 146–166.

Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge, MA, and London: Harvard University Press.

Roemer, J. E. and A. Trannoy (2015). Chapter 4 - Equality of Opportunity. In A. B. Atkinson and F. Bourguignon (Eds.), *Handbook of Income Distribution*, Volume 2 of *Handbook of Income Distribution*, pp. 217–300.

Roemer, J. E. and A. Trannoy (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature 54*(4), 1288–1332.

Russell, S. J. and P. Norvig (2016). *Artificial intelligence: a modern approach*. Pearson.

Scanlon, T. (2018). *Why Does Inequality Matter?* Oxford University Press.

Scutari, M., C. E. Graafland, and J. M. Gutiérrez (2019). Who learns better Bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning 115*, 235–253.

Sen, A. (1980). *Equality of What?* Cambridge University Press.

Shorrocks, A. (1980). The Class of Additively Decomposable Inequality Measures. *Econometrica 48(3)*, 613–625.

Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine learning 65*(1), 31–78.

A Additional figures and tables

Table A.1: Unconditional distribution of father education in Italy

Variable		**Probability**
Education of father	*Low*	0.70
	Medium	0.23
	High	0.07

Note: The table reports the unconditional distribution of the categorical variable *Education of Father* (*Low* for primary education, *Medium* for secondary education, *High* for tertiary education) for Italy, using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. *Source*: Authors's elaborations on EU-SILC data 2019.

Table A.2: Conditional distribution of paternal occupation in Italy

		Occupation of father		
		Blue-collar	*White-collar*	*Manager*
Education of father	*Low*	0.37	0.54	0.09
	Medium	0.14	0.46	0.40
	High	0.03	0.08	0.89

Note: The table reports the distribution of the categorical variable *Occupation of father* (Blue-collar, White-collar, Manager) conditional on the categorical variable *Education of father* (*Low* for primary education, *Medium* for secondary education, *High* for tertiary education) for Italy, using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. Father's occupation is obtained by combining the activity status and the ISCO-8 classification reported in EU-SILC. Blue-collars also include unemployed and inactive fathers. *Source*: Authors's elaborations on EU-SILC data 2019.

Table A.3: Conditional distribution of financial distress at age 14 in Italy

		Financial distress	
		No	*Yes*
Occupation of father	*Blue-collar*	0.73	0.27
	White-collar	0.82	0.18
	Manager	0.95	0.05

Note: The table reports the distribution of the categorical variable Occupation of father (Blue-collar, White-collar, Manager) conditional on the indicator for bad financial situation when the respondent was at age 14 (*Financial distress*) for Italy in 2019, using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. Father's occupation is obtained by combining the activity status and the ISCO-8 classification reported in EU-SILC. Blue-collars also include unemployed and inactive fathers. *Financial distress* is obtained by recoding the variable on the perceived financial situation in the household at age 14 provided by EU-SILC. *Source*: Authors's elaborations on EU-SILC data 2019.

Table A.4: Conditional distribution of market income in Italy (2018)

		Labour earnings (€)	
		Mean	**SD**
Occupation of father	*Blue-collar*	23,288.52	23,373.86
	White-collar	27,279.99	25,699.15
	Manager	35,131.51	35,371.61

Note: The table reports the mean and standard deviation of *Labour earnings* in 2018 in Italy conditional on the categorical variable *Occupation of father* (Blue-collar, White-collar, Manager), using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. Labour earnings in Euros is measured at the individual level in the year, coming from working activities, either as employee or as self-employed, gross of social insurance contributions and taxes. Father's occupation is obtained by combining the activity status and the ISCO-8 classification reported in EU-SILC. Blue-collars also include unemployed and inactive fathers. *Source*: Authors's elaborations on EU-SILC data 2019.

Table A.5: Descriptive Statistics

Variable	Poland	Germany	Italy	France	Sweden
Labour earnings (€)					
Mean	10,161.96	36,094.72	27,672.84	39,766.58	42,045.26
SD	9,661.16	33,544.58	27,908.31	33,373.01	30,165.18
Cohort of birth					
1959-1969	0.32	0.46	0.39	0.36	0.36
1970-1979	0.29	0.29	0.32	0.31	0.31
1980-1993	0.39	0.25	0.28	0.33	0.34
Sex					
Women	0.50	0.54	0.51	0.52	0.47
Men	0.50	0.46	0.49	0.48	0.53
Country of birth					
Local	0.80	0.92	0.87	0.90	0.83
EU	0.00	0.00	0.04	0.03	0.05
Extra-EU	0.00	0.08	0.08	0.08	0.12
Other/Unknown	0.19	0.00	0.00	0.00	0.00
Education of father					
Low	0.25	0.09	0.67	0.64	0.33
Medium	0.44	0.47	0.22	0.09	0.27
High	0.06	0.30	0.07	0.14	0.29
Other/Unknown	0.26	0.15	0.04	0.13	0.11
Education of mother					
Low	0.27	0.09	0.73	0.68	0.32
Medium	0.41	0.51	0.20	0.11	0.30
High	0.06	0.12	0.04	0.11	0.29
Other/Unknown	0.26	0.27	0.04	0.10	0.09
Occupation of father					
Unemployed/Inactive	0.02	0.02	0.03	0.04	0.03
Low-skilled	0.18	0.16	0.25	0.30	0.13
Medium-skilled	0.42	0.39	0.47	0.39	0.35
High-skilled	0.12	0.36	0.21	0.28	0.39
Other/Unknown	0.25	0.07	0.04	0.00	0.10
Occupation of mother					
Unemployed/Inactive	0.14	0.32	0.60	0.42	0.16
Low-skilled	0.09	0.12	0.09	0.13	0.08
Medium-skilled	0.38	0.28	0.18	0.31	0.40
High-skilled	0.14	0.19	0.10	0.14	0.30
Other/Unknown	0.25	0.09	0.03	0.00	0.07
Household type					
Both parents	0.73	0.93	0.98	0.92	0.77
One parent	0.01	0.00	0.01	0.08	0.05
No parents	0.00	0.00	0.00	0.01	0.00
Other/Unknown	0.25	0.07	0.01	0.00	0.18
Number of siblings					
No or one sibling	0.45	0.71	0.78	0.64	0.61
Two or three siblings	0.24	0.19	0.18	0.28	0.29
Four or more siblings	0.05	0.02	0.02	0.08	0.04
Other/Unknown	0.26	0.07	0.01	0.00	0.06
Urbanization					
City	0.13	0.22	0.23	0.17	0.26
Town/suburb	0.16	0.29	0.45	0.33	0.43
Rural area	0.46	0.43	0.32	0.50	0.24
Other/Unknown	0.25	0.07	0.01	0.00	0.08
Financial distress					
Yes	0.15	0.16	0.18	0.24	0.17
No	0.58	0.76	0.80	0.74	0.77
Other/Unknown	0.26	0.09	0.02	0.02	0.07
Number of observations	17,502	7,222	15,830	7,280	2,234
Weighted population	14,702,540	26,839,698	23,670,086	18,441,987	2,013,365

Note: The table reports summary statistics (mean and standard deviation for labour earnings, and proportion in the sample for the other variables) for each country using the appropriate sample weights. The sample includes respondents aged between 25 and 59 who do not declare to be students, retired, or unable to work due to disability. *Source*: Authors's elaborations on EU-SILC data 2019.