Credible causal inference beyond toy models (Or: Building DAGs bottom-up for better empirical research)

Pablo Geraldo, UCLA Sociology and Statistics

Motivation

The social sciences are experimenting what some authors have called a "credibility revolution" (Angrist and Pischke, 2010), the rise of "causal empiricism" (Samii, 2016), or simply a "causal revolution" (Pearl and MacKenzie, 2018).

The enormous progress in the last decades is associated with the development of two frameworks that allow researchers to transparently handle causal questions: **Potential Outcomes** (Neyman-Rubin) and the **Structural Causal Model** (Wright-Pearl).

Both languages are formally equivalent, but not equally expressive.

The problem: Different disciplines and research communities have adopted each language separately, producing diverging research practices:

- In the PO tradition, emphasis on quasi-experiments and so-called "research templates": Keep design and analysis simple, calls for *design-based* causal inference.
- In the SCM tradition, emphasis on observational studies and understanding the entire DGP: Dealing with complex processes using DAGs, recognition of causal inference as *model-based*.

Where we agree: Causal inference require untestable assumptions ("No causes in, no causes out")

Where we disagree: How to encode and assess the "causes in"? What constitutes *credible* causal inference? How much "modeling" does it require?

Diagnosis: The choice between *model-free* as opposed to *model-based* causal inference is a false dilemma, diverting us from addressing the mostly harmful practice of *model-blind* research

- Empirical research do not univocally map into research templates ("identification strategies")
- Quasi-experiments do not totally bypass the need to model the process under study
- Observational studies do not require to model the entire data generating process
- Causal graphical models can help to assess the strength of evidence *as it is*

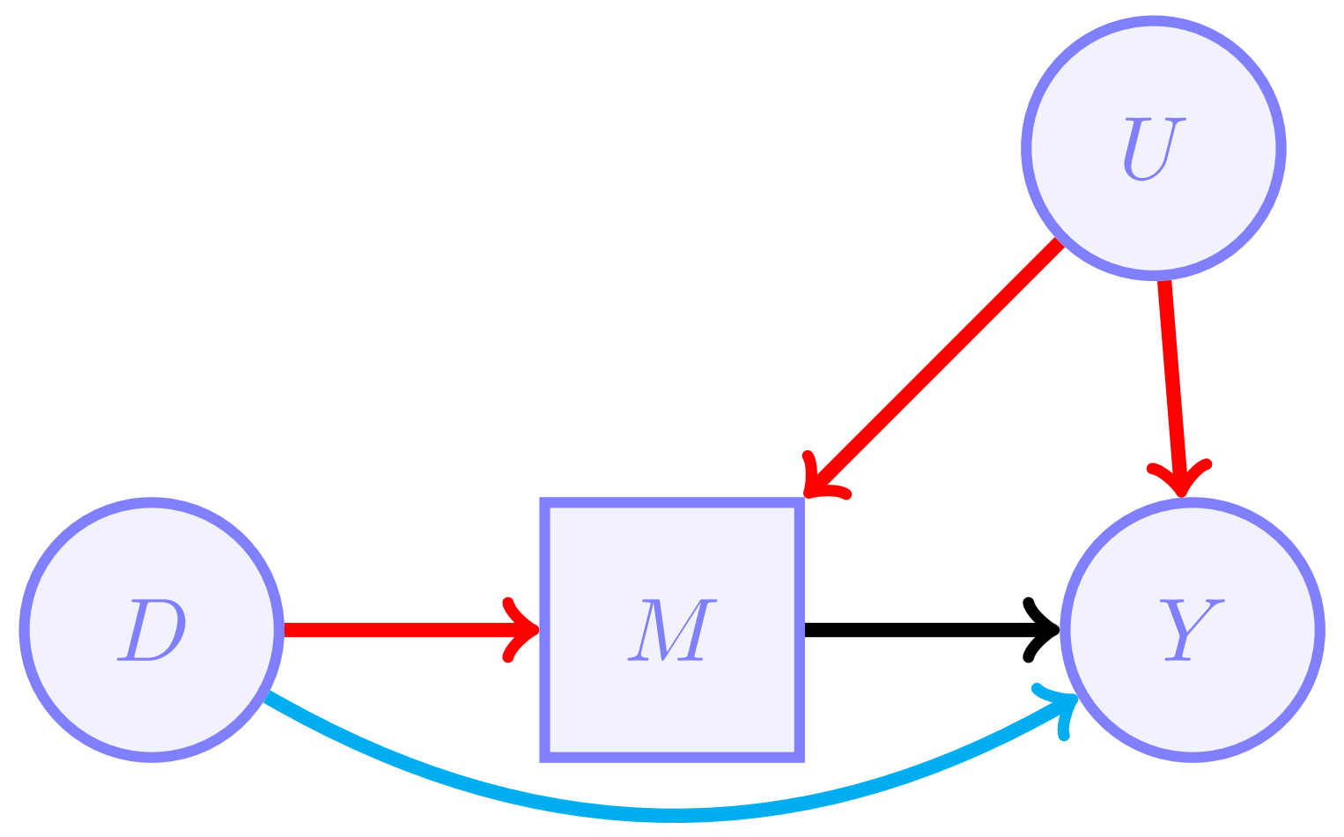
To DAG or not to DAG?

Causal inference without DAGs: Somewhat ironically, graphical models in causal inference have been advised against on two different grounds: they are either trivial or dangerous.

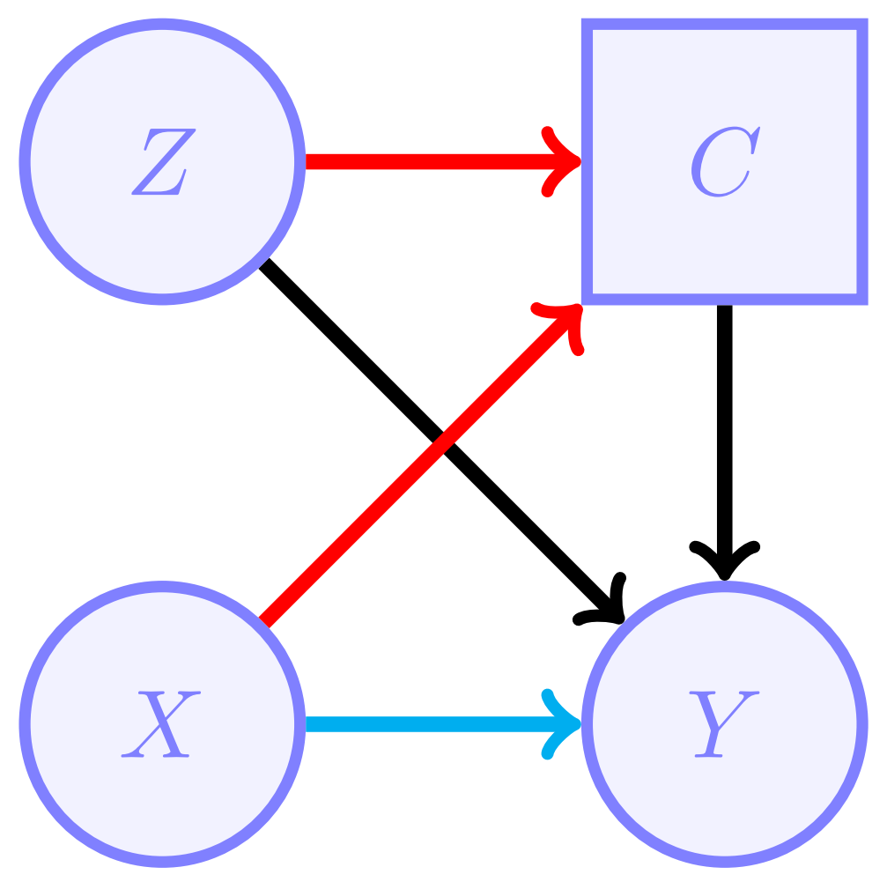
"Credible empirical research is too simple for DAGs to be useful; settings where DAGs can prove useful are too complex to be credible"

Causal inference with DAGs: Proponents of DAGs has emphasized their utility in different stages of the research process:

- During training, DAGs are used as *graphical scaffoldings*, intuition-building devices that can be discarded when the solid building (internalizing identification strategies) is finished
- Graphical models can also be used as abstractions, to build *negative templates*: representative failures of identification attempts. Some examples:



Research question: Can we use police administrative records to estimate racial discrimination in policing? **Diagnosis:** Target quantity unclear, collider bias estimating direct effect of race even within selected sample of stops. **Proposal:** The authors propose a bias-correction method, a bounding procedure, and a design that is not affected by the identified problems. **References:** Knox et al. (2020) discussing Antonovics and Knight (2009), Fryer (2018, 2019), Johnson (2019), and Ridgeway (2006)



Research question: Is intergenerational mobility higher among college graduates than non graduates? **Diagnosis:** Current estimates suffer from collider bias (conditioning on C, college completion), thus distorting estimates of intergenerational mobility ($X \to Y$). **Proposal:** Residual balancing, a method to break the link between Z (joint determinants of C and Y) and college completion (C) before conditioning on C, for unbiased estimation of $X \to Y$. **References:** Zhou (2019) discussing Hout (1984, 1988) Breen (2010) Chetty et al. (2017) Pfeffer and Hertel (2015), and Torche (2011)

Building DAGs bottom-up

A step-by-step guide of building DAGs "bottom-up", starting from the assumptions hidden in a research design.

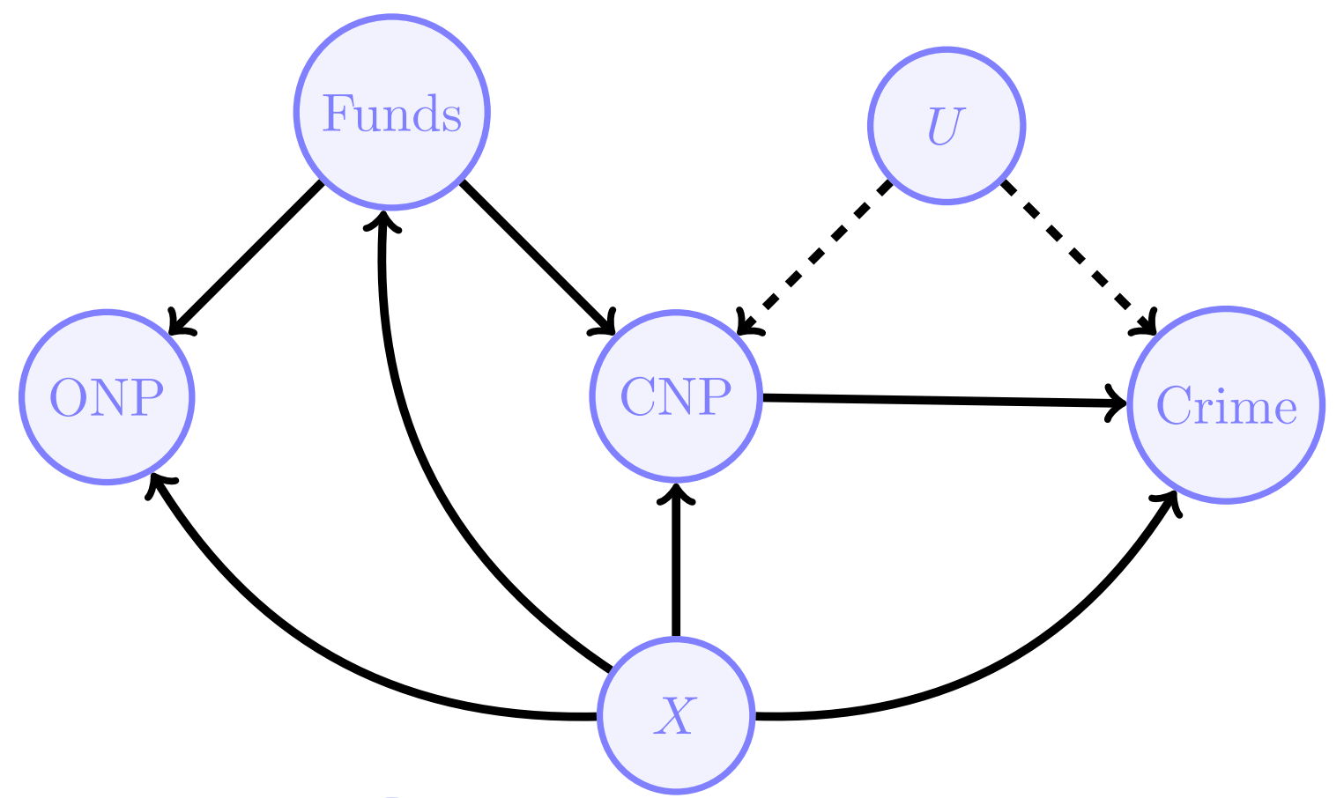
1. From the reading of a particular study, construct the DAG inductively:
   What variables are being included in the analysis?
   What relationships between variables are being assumed?
2. If the exercise of reverse-engineering the implied model provides ambiguous results:
   Build a set of models compatible to the information provided by the authors.
   (This could be an equivalent class or a group of incompatible models)
3. Evaluate the identification argument:
   Is the quantity of interest identified under the authors' own model?
4. Assess the credibility of the assumptions encoded in the resulting model:
   a. Assess the credibility "internally":
   is there any missing relationship between the variables already included?
   Does this change the conclusions?
   b. Asses the credibility "externally":
   is there any missing variable that should have been included?
   Does this change the conclusions?
5. If necessary, build an alternative diagram:
   Taking into account the results from the model criticism exercise mentioned above.
6. Systematically derive testable implications of the competing models:
   Which variables should be (conditionally) independent or "balanced" if the model is true?
7. Test the compatibility of the model to the observed data and results:
   This includes conducting sensitivity analysis and falsification tests.
8. Update and repeat.

Table 1. Proposed steps to build DAGs inductively

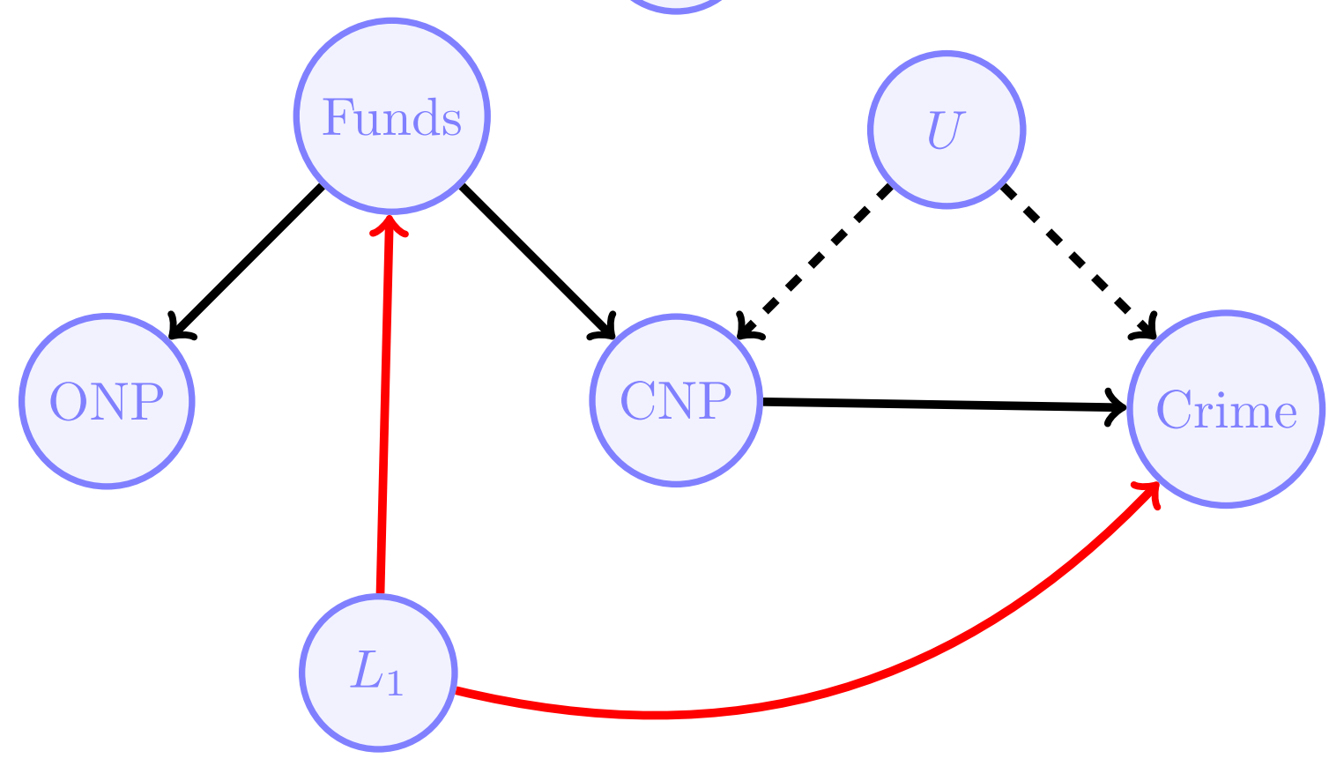
Depending on the circumstances, one might need to follow a subset of steps:

- **For researchers** when conducting an observational study, steps 1-7 are ideal to provide a transparent account of the empirical analysis
- **For reviewers** when assessing the evidence of a research piece, steps 1-4, and ideally 5, would promote fruitful exchanges with authors
- **For the scientific community** as a whole, step 8 is core of developing the literature.

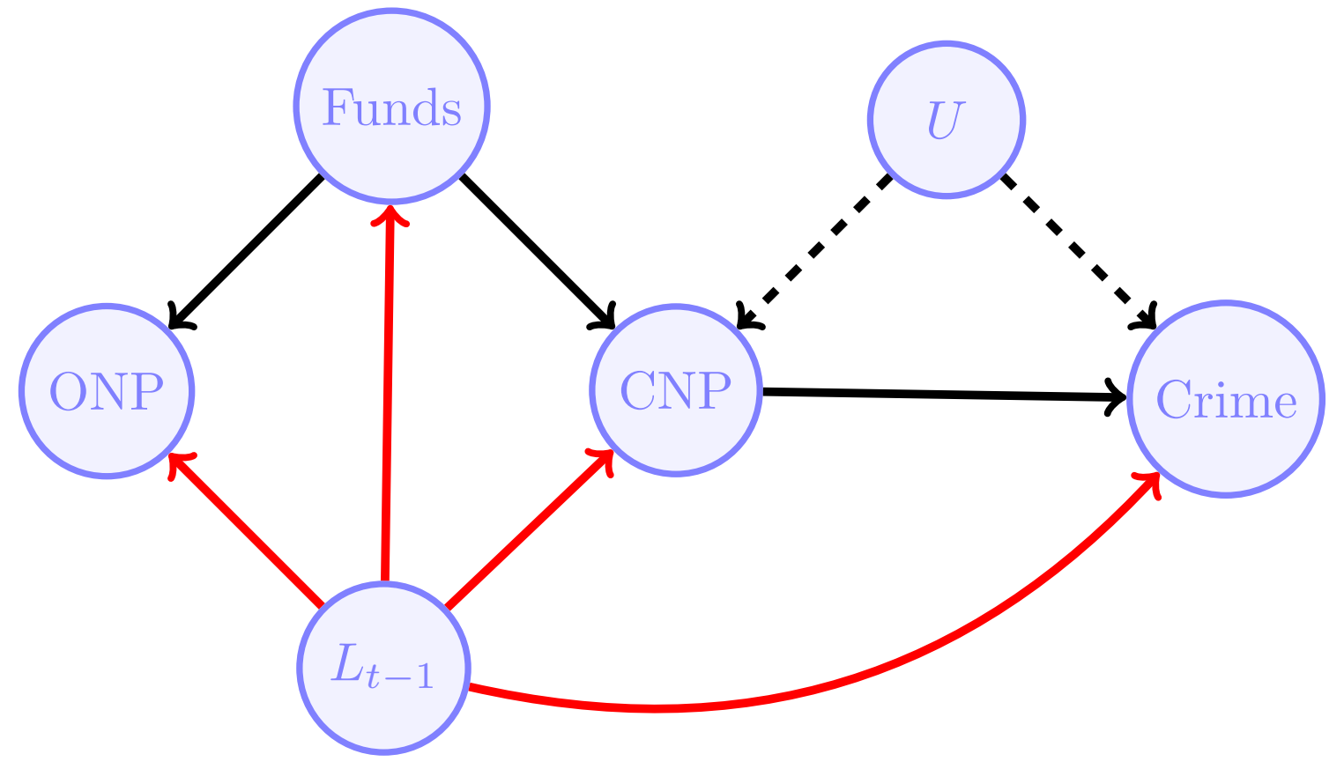
Example 1: Sharkey et al (2017)



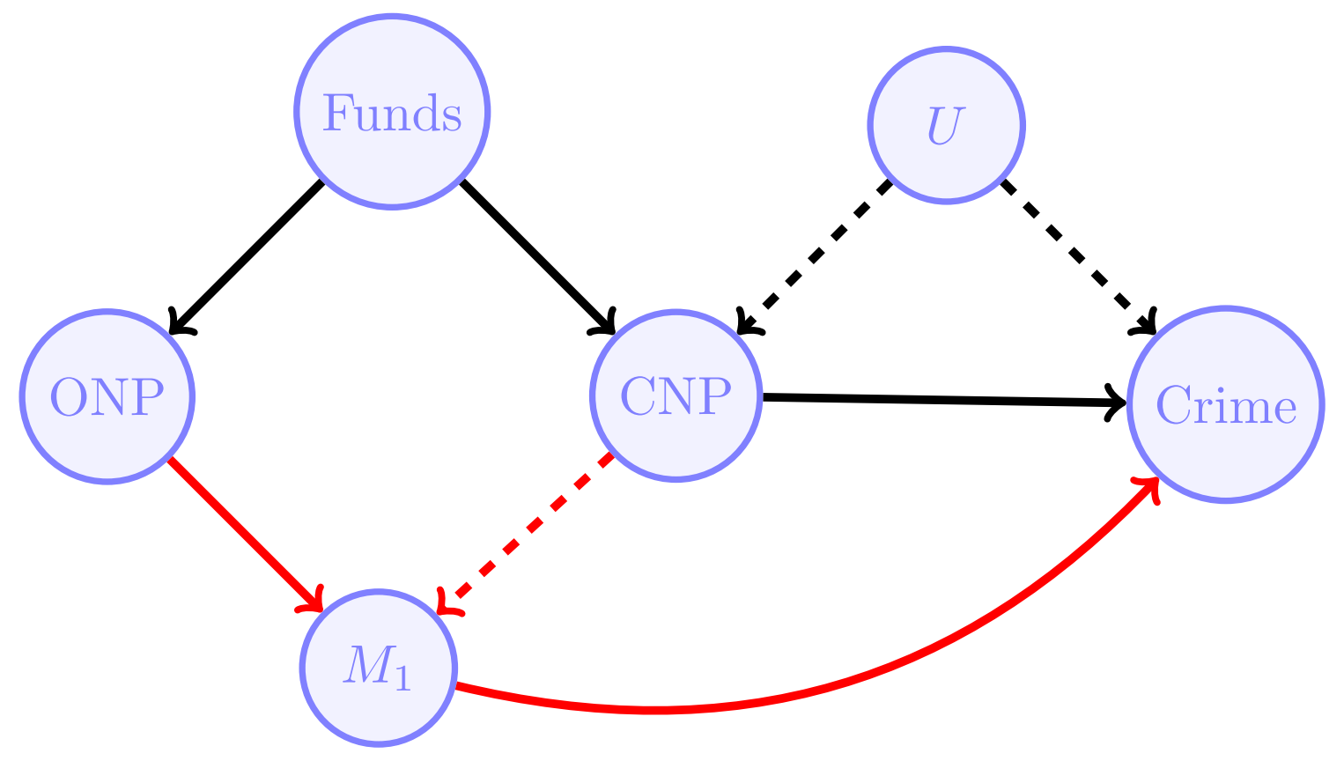
Instrumental Variable setting from Sharkey et al. (2017): The variables are: availability of funding (Funds), community nonprofits (CNP), other nonprofits (ONP), registered violent crime (Crime), and unobserved common factors between CNP and Crime (U). Control variables in X are: population density, ethnic composition, educational composition, sex by age composition, immigration percentage, unemployment, and occupational composition.



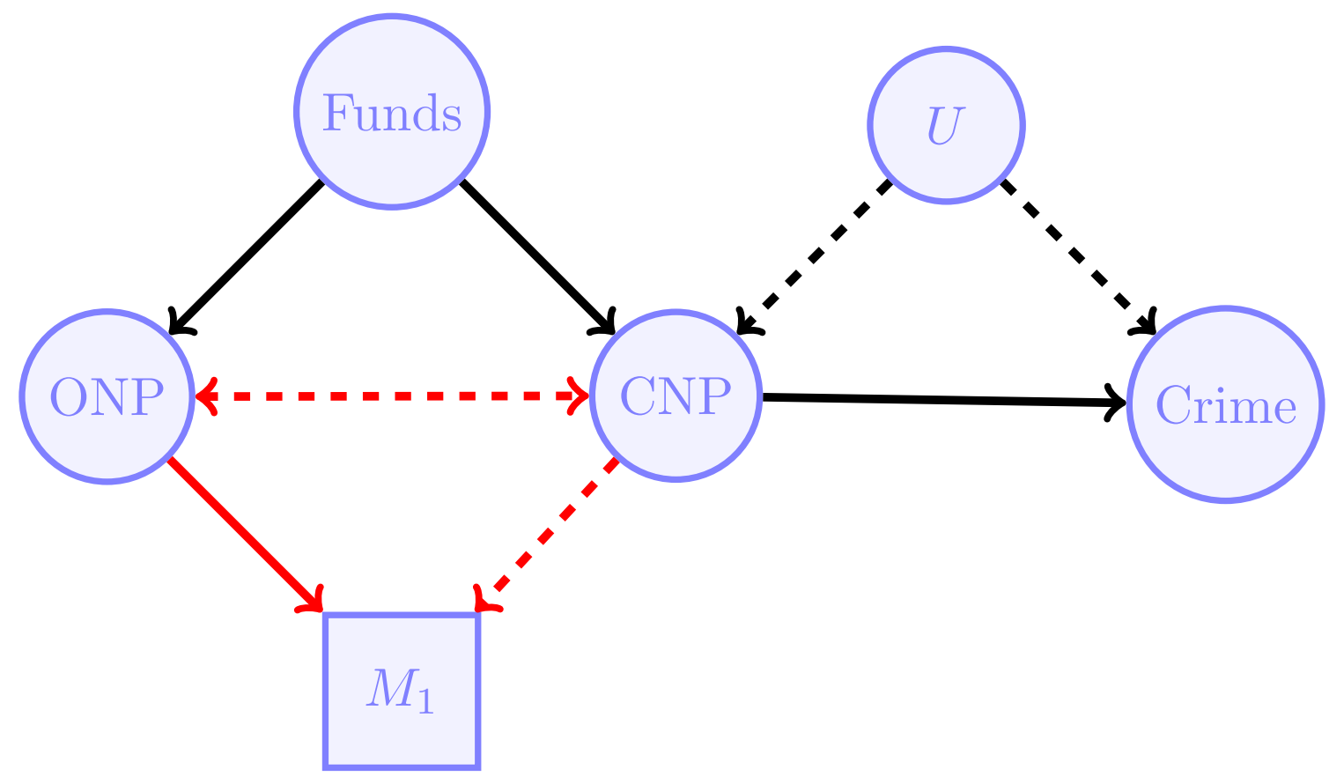
(A) The (unobserved) instrument shares a common cause with the outcome. For example, political orientation of the local government could affect funds available for community organization and policing programs simultaneously.



(B) The instrument, the surrogate instrument and the treatment share a common cause with the outcome. For example, previous levels of community involvement could affect the availability of funds and have a direct effect on crime.

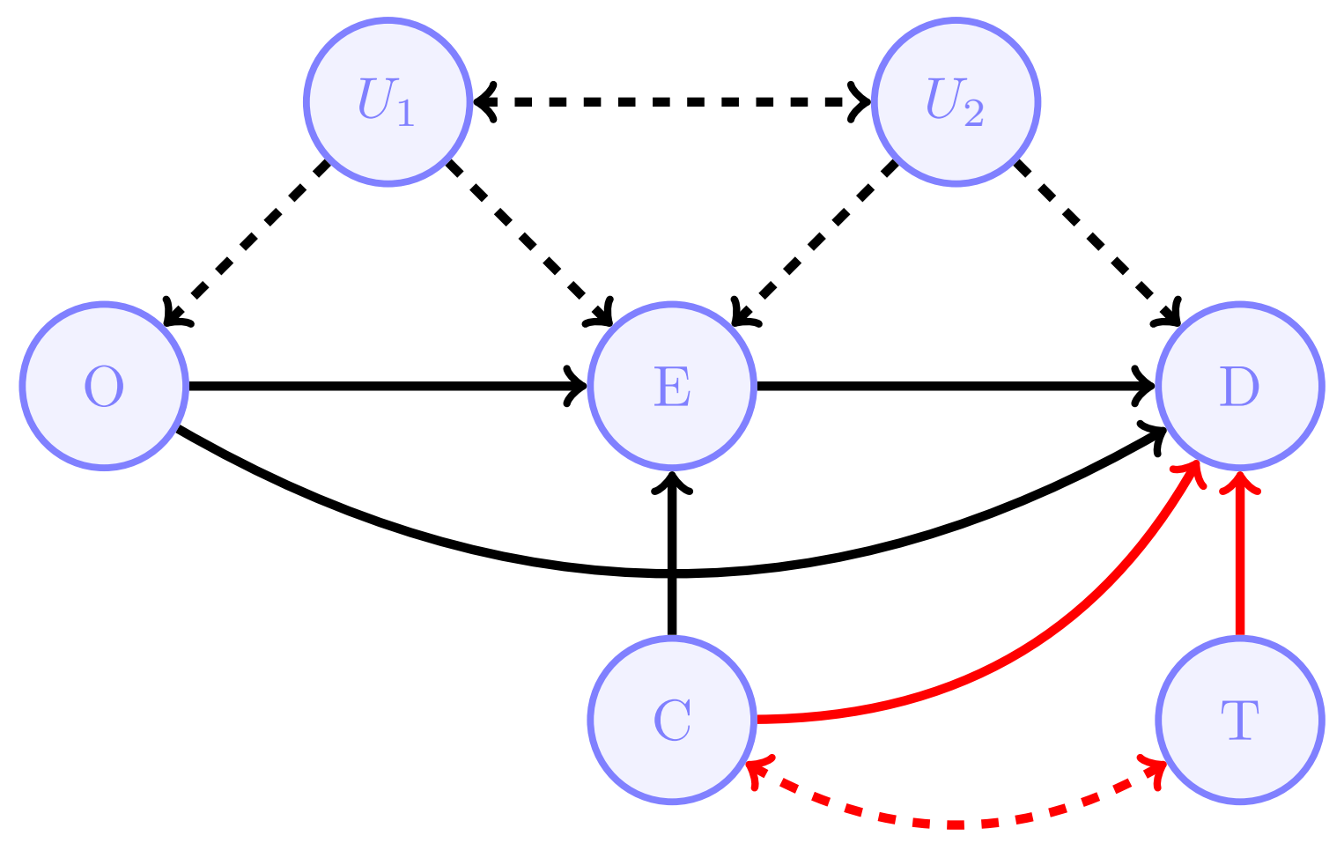


(C) An hypothetical mechanism M_1 that mediates the (indirect) effect of Other Nonprofits on Crime. For example, building a local community and increasing social capital.

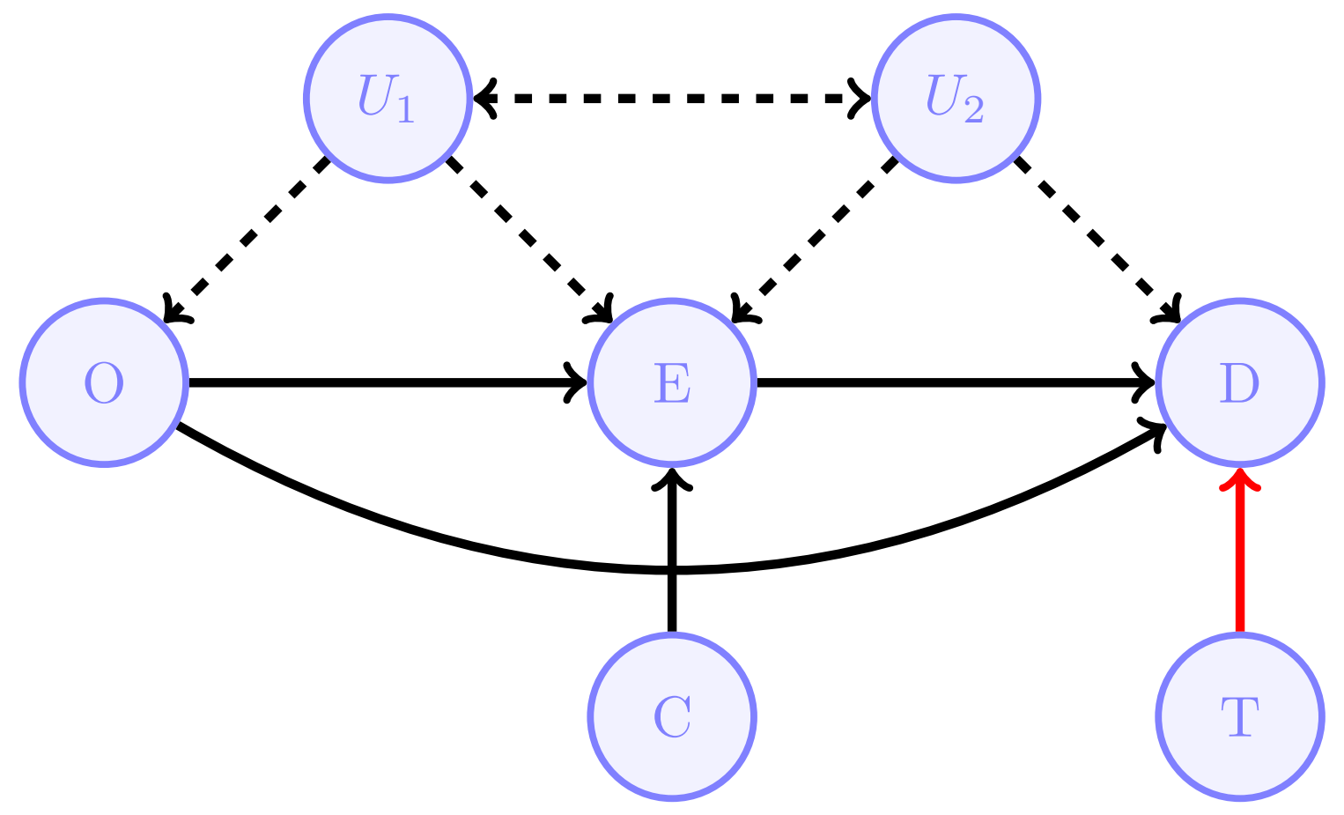


(D) If the mechanism is shared between all Community Nonprofit and Other Nonprofits, conditioning on the mechanism do not solve the problem, inducing collider bias.

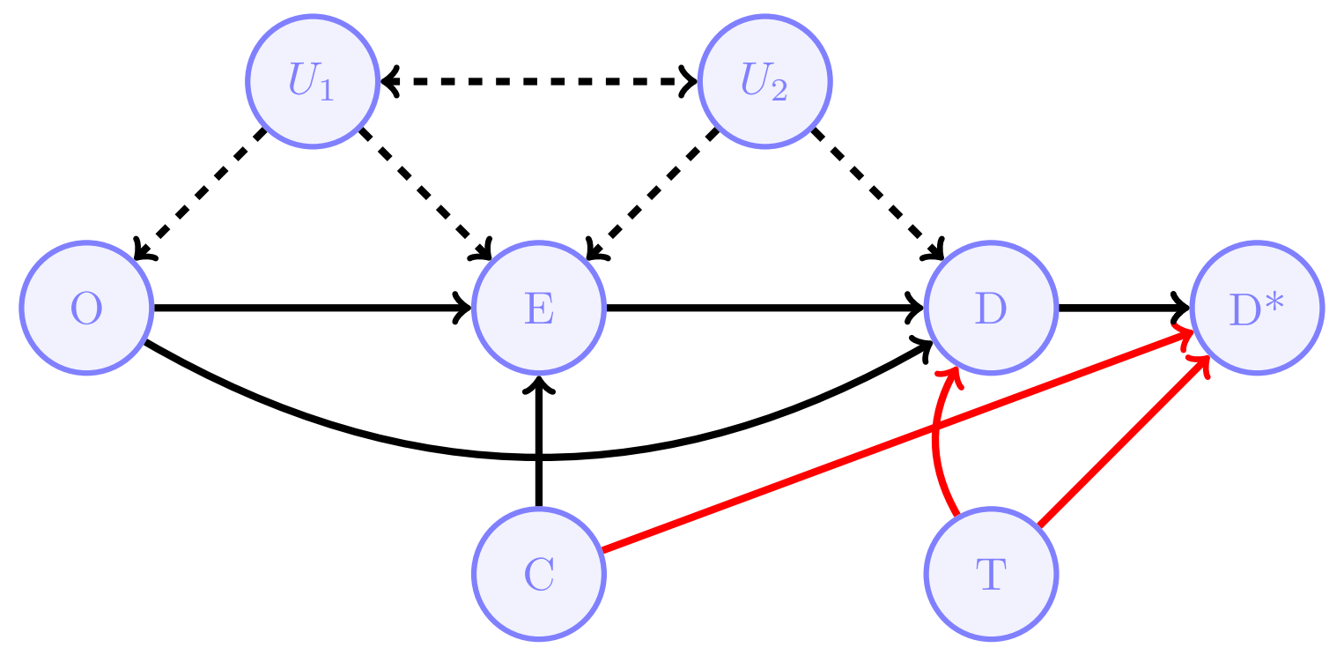
Example 2: Rauscher (2016)



RDD on time setting from Rauscher (2016): Data generating graph: Cohort (C) is associated with educational attainment (E) through educational expansion, with attained socioeconomic status (D) through labor market experience, and other social and economic time trends (T) that affecting mobility.



Limiting graph: In a narrow window around the cohort cutoff for exposure (C), it is expected that cohort becomes unrelated with social destination (D) and mobility time trends (T), except for its effect on education (E). Cohort becomes a valid IV.



Threat: If social destination (D) is measured at different points in time (D*) depending on the cohort, then cohort will not be a valid instrument anymore.

Conclusions

- Actual empirical research do not perfectly maps into research templates (canonical identification strategies)
- Classifying research according to an ideal template can be misleading, obscuring causal assumptions and signaling credibility instead of putting them forward for critical examination
- Using DAGs to express schematic research templates falls short in exploiting the expressive power of graphical models
- DAGs can help to understand how a particular setting deviates from the ideal template, improving research design and analysis
- When using DAGs, one doesn't need to commit *ex ante* to a full model of the DGP; these models naturally grow in every empirical application and can be reconstructed inductively

References

Pablo Geraldo, "DAGs beyond toy models: Using causal graphs for better empirical research", Draft

Emily Rauscher. Does Educational Equality Increase Mobility? Exploiting Nineteenth-Century U.S. Compulsory Schooling Laws. American Journal of Sociology, 121(6):1697–1761, May 2016

Patrick Sharkey, Gerard Torrats-Espinosa, and Delaram Takyar. Community and the Crime Decline: The Causal Effect of Local Nonprofits on Violent Crime. American Sociological Review, 82(6):1214–1240, December 2017