Title: Quantum Foundations SeminarBayesian learning of Causal Structure and Mechanisms with GFlowNets and Variational Bayes

Speakers: Mizu Nishikawa-Toomey

Series: Quantum Foundations

Date: May 28, 2024 - 2:00 PM

URL: https://pirsa.org/24050094

Abstract: Bayesian causal structure learning aims to learn a posterior distribution over directed acyclic graphs (DAGs), and the mechanisms that define the relationship between parent and child variables. By taking a Bayesian approach, it is possible to reason about the uncertainty of the causal model. The notion of modelling the uncertainty over models is particularly crucial for causal structure learning since the model could be unidentifiable when given only a finite amount of observational data. In this paper, we introduce a novel method to jointly learn the structure and mechanisms of the causal model using Variational Bayes, which we call Variational Bayes-DAG-GFlowNet (VBG). We extend the method of Bayesian causal structure learning using GFlowNets to learn not only the posterior distribution over the structure, but also the parameters of a linear-Gaussian model. Our results on simulated data suggest that VBG is competitive against several baselines in modelling the posterior over DAGs and mechanisms, while offering several advantages over existing methods, including the guarantee to sample acyclic graphs, and the flexibility to generalize to non-linear causal mechanisms.

Zoom link

Bayesian modelling of structural causal models using GFlowNets

Mizu Nishikawa-Toomey, Tristan Deleu, Jithendaraa Subramanian,
Yoshua Bengio & Laurent Charlin

Content

- (Very quick) Intro to Bayesian machine learning
- Causal structure learning
- GFlowNets
- Causal structure learning using GFlowNets

Bayesian machine learning VS non-Bayesian

Machine learning is about creating useful models of the world.
Most of machine learning is about given data, D find the model M that maximises the likelihood of the data.

$$\arg\max_{M} P(D \mid M)$$

What is the problem with this approach?
Given a finite amount of data, we do not know which model is closest to the true data generating process, and many models will have the same likelihood.

This is one argument of taking a Bayesian approach to machine learning.

Bayesian machine learning

Maximum likelihood machine learning: $\arg\max_M P(D \mid M)$

In Bayesian machine learning, we find:

$$P(M \mid D)$$

A whole set of models that describe the given data. How do we do this?

Using Bayes rule:

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)}$$

This is actually pretty hard to do because:

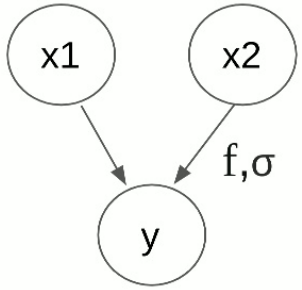
$$P(D) = \int P(D \mid M)P(M)dM$$

Causal modelling

Structural causal models consists of d structural assignments [1] :

$$X_j := f_j\left(\mathbf{PA}_j, N_j\right), \quad j = 1, \ldots, d$$

$$M = (G, f, \sigma^2)$$



Causal structure learning has the goal of finding the graph **G**, given some assumptions on the observed data.

[1] Peters, J.; Janzing, D. & Schölkopf, B. (2017), Elements of Causal Inference: Foundations and Learning Algorithms , MIT Press , Cambridge, MA .

Causal structure learning

- Causal structure learning algorithms broadly go in to two categories: constraint based (eg. PC [2]) and score based (eg GES [3]).
- Often in score based approaches, an assumption is made for the functional relations so we can formulate a likelihood for the data for the score.

Number of realisations of that same graph (n)

Number of nodes in graph (k)

	node1	node2	node3	node4	node5
0	0.683485	0.952184	0.904999	0.496995	0.018659
1	0.575472	0.044466	0.556824	0.484083	0.981189
2	0.730692	0.011476	0.784822	0.040864	0.670788
3	0.102671	0.395069	0.836394	0.976337	0.236687
4	0.294205	0.912545	0.864391	0.998549	0.928131
5	0.397266	0.905975	0.609466	0.556205	0.640405
6	0.460048	0.078794	0.508832	0.172824	0.781167
7	0.150959	0.652586	0.596166	0.901904	0.441611
8	0.993209	0.537675	0.258738	0.821872	0.970752
9	0.992899	0.400410	0.019881	0.269530	0.046813
10	0.904563	0.328384	0.683180	0.353318	0.276211
11	0.396885	0.411343	0.053515	0.594002	0.071475
12	0.973078	0.360488	0.190598	0.541303	0.756543
13	0.956749	0.510627	0.559052	0.435750	0.486829
14	0.239131	0.005482	0.455422	0.559531	0.145727

$$P(X \mid G, \theta) = \prod_{n=1}^{N} \prod_{k=1}^{K} P(X_k^{(n)} \mid \mathrm{Pa}_G(X_k^{(n)}), \theta_k)$$

[2] Peter Spirtes, Clark Glymour, and Richard Scheines.Causation, Prediction, and Search,volume 81. 01 1993
[3] D. M. Chickering. Optimal structure identification with greedy search.Journal of Machine Learning Research, 3:507–554, 2002b

Markov equivalence classes

- Using observational data, we can only identify Markov equivalence classes of DAGs (one of these boxes).
- Two DAGs are (likelihood) equivalent iff the undirected skeleton and the V-structure are the same. [1]
- It important that our DAG inference algorithm suggests multiple candidates of graphs by design, rather than just one.

$$P(M \mid D)$$

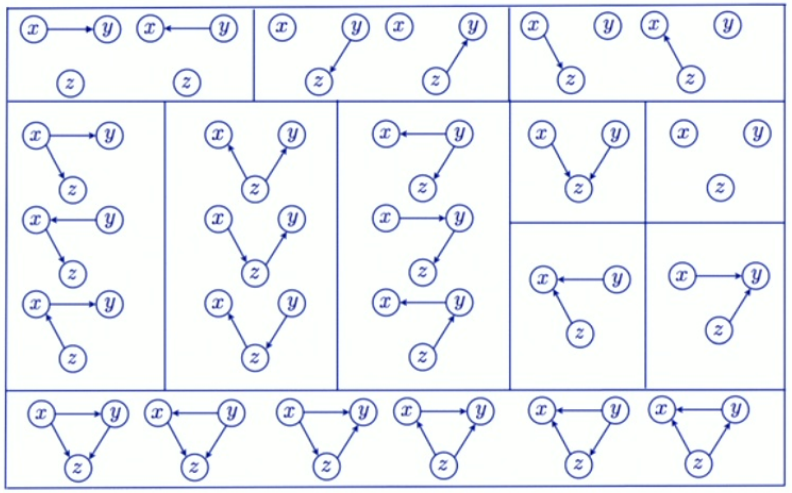
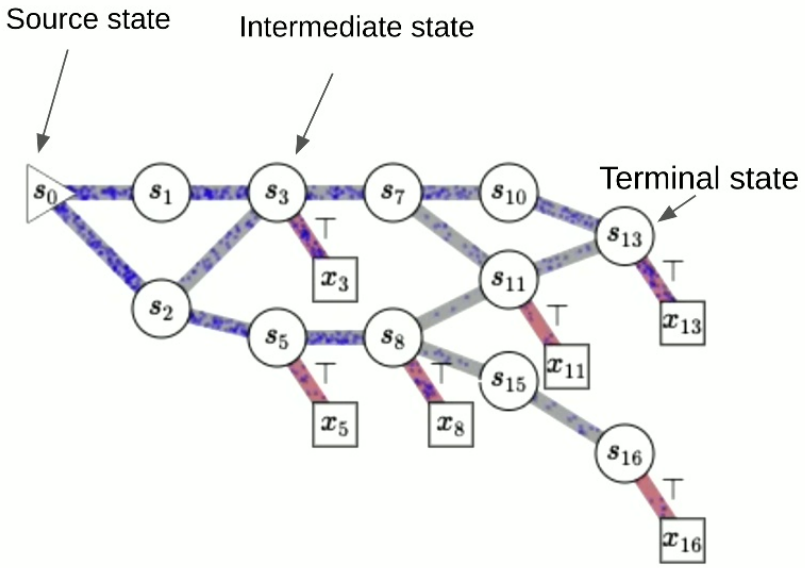


Image :Introduction to the foundations of causal discovery - Scientific Figure on ResearchGate. [accessed 28 May, 2024]

[1] Peters, J.; Janzing, D. & Schölkopf, B. (2017), Elements of Causal Inference: Foundations and Learning Algorithms , MIT Press , Cambridge, MA .

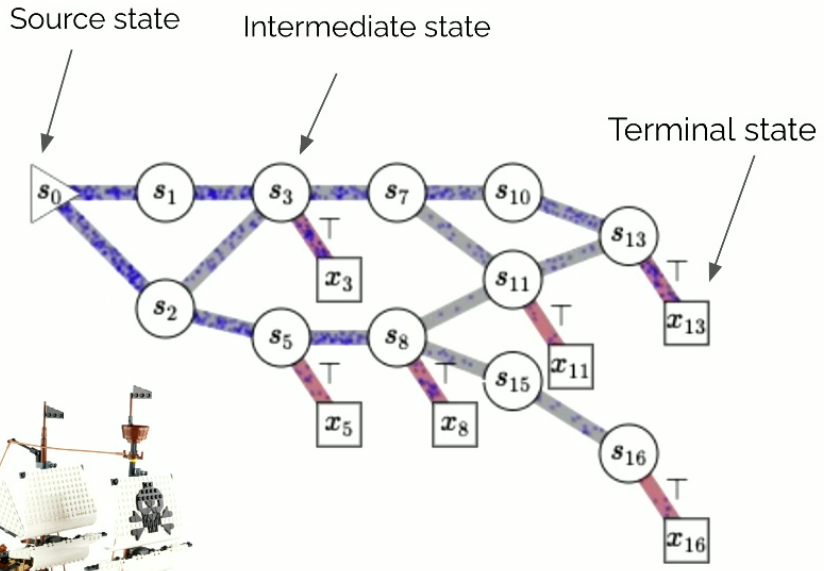
GFlowNets



Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. Flow network based generative models for non-iterative diverse candidate generation. NeurIPS'2021, arXiv:2106.04399, 2021.

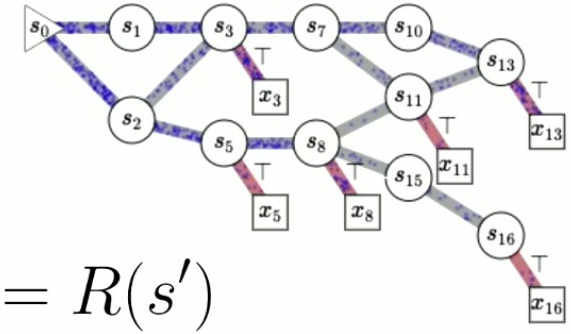
States of a GFlowNet

- There are three types of states. Source, intermediate and terminal. Intermediate states have 0 reward, terminal states have a positive reward, a single source state has a reward equal to the sum of all rewards of terminal states. Only terminal states are complete objects **x,** subject to certain conditions.
- Each state **x** in the GFlowNet has an associated reward **R(x)**. Training data consists of **x, R(x)** consists of states and corresponding reward for that states.



Bengio, M. Jain, M. Korablyov, D. Precup, and Y. Bengio. Flow network based generative models for non-iterative diverse candidate generation. NeurIPS'2021, arXiv:2106.04399, 2021.

Between the states of a GFlowNet



flow consistency:

$$\sum_{s,a:T(s,a)=s'} F(s,a) - \sum_{a' \in A(s')} F(s',a') = R(s')$$

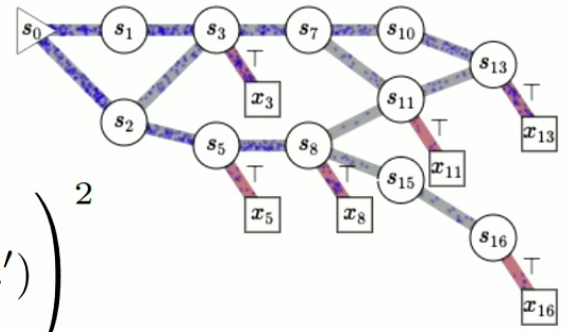
$$\pi(a|s) = \frac{F(s,a)}{\sum_{a'} F(s,a')}$$

The Loss function

flow consistency:

$$\sum_{s,a:T(s,a)=s'} F(s,a) - \sum_{a' \in A(s')} F(s',a') = R(s')$$

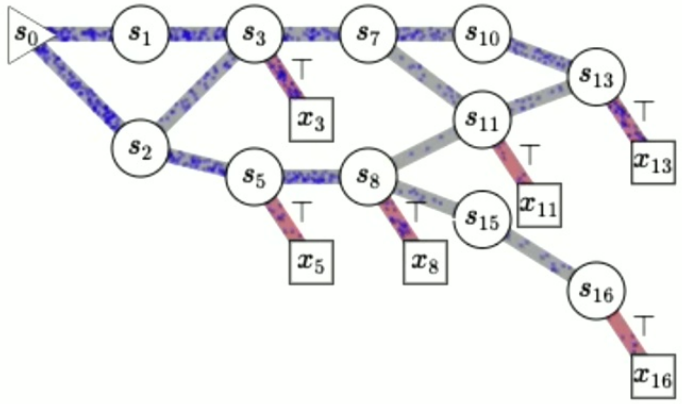
Loss function is based around the above criteria for a single trajectory:



$$\tilde{\mathcal{L}}_\theta(\tau) = \sum_{s' \in \tau \neq s_0} \left(\sum_{s,a:T(s,a)=s'} F_\theta(s,a) - R(s') - \sum_{a' \in \mathcal{A}(s')} F_\theta(s',a') \right)^2$$

Sampling from a GFlowNet

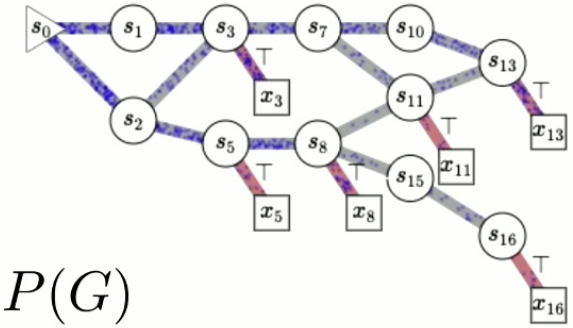
$$\pi(a|s) = \frac{F(s,a)}{\sum_{a'} F(s,a')}$$



Start at the source node and take transitions according to the policy until we arrive at a terminal state.

How can we use GFlowNets for Bayesian causal structure learning?

Specify a reward function for graphs.
Then train the GFlowNet on **(R(G), G)** pairs.
What is the reward? The un-normalised posterior.



$$P(G \mid D) = \frac{P(D \mid G)P(G)}{P(D)} \propto P(D \mid G)P(G)$$

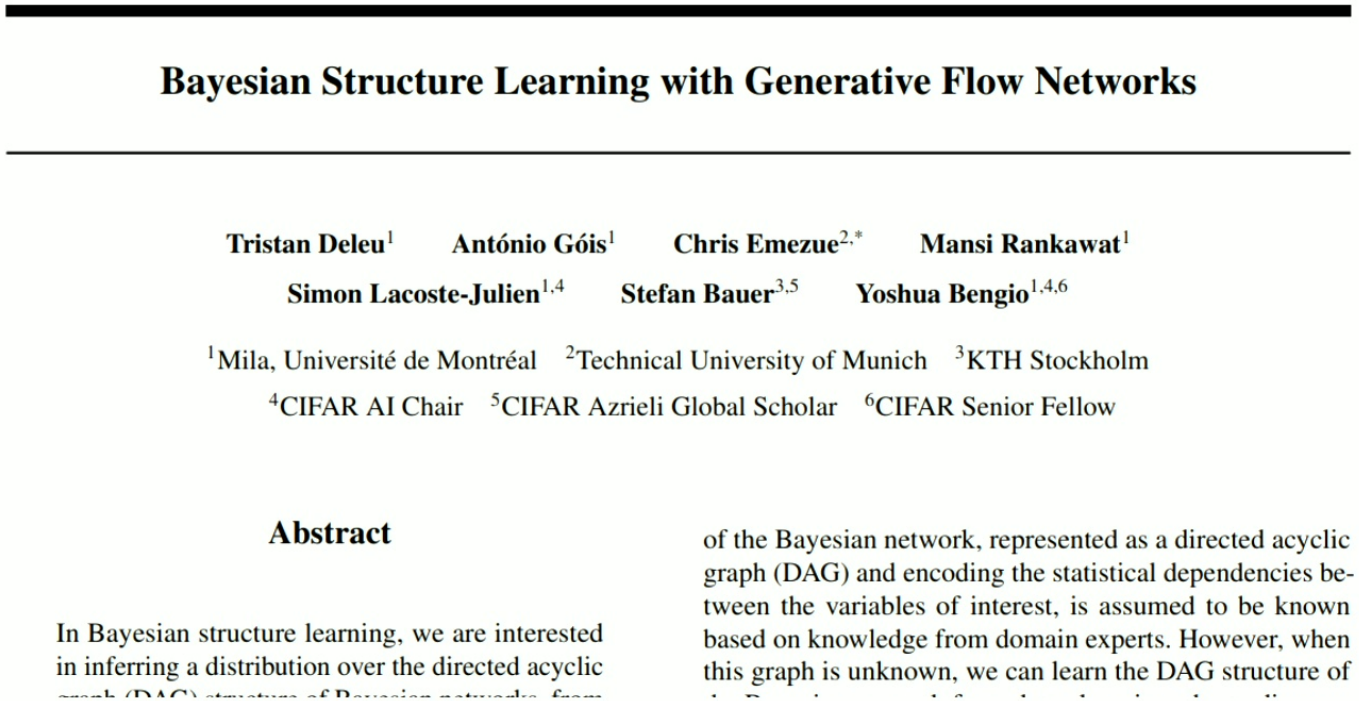
$$R(G) = P(D \mid G)P(G)$$

Then we can sample graphs proportional to the posterior.
However we have made a number of assumptions to calculate the likelihood, P(D|G) from the data.

Assumptions:

- A model for the likelihood P(D|G)
- No unobserved confounders
- Faithfulness
- Acyclicity

GFlowNets for causal structure learning (the first paper)



Bayesian Structure Learning with Generative Flow Networks

Tristan Deleu[1] **António Góis**[1] **Chris Emezue**[2,*] **Mansi Rankawat**[1]

Simon Lacoste-Julien[1,4] **Stefan Bauer**[3,5] **Yoshua Bengio**[1,4,6]

[1]Mila, Université de Montréal [2]Technical University of Munich [3]KTH Stockholm

[4]CIFAR AI Chair [5]CIFAR Azrieli Global Scholar [6]CIFAR Senior Fellow

28 Jun 2022

Abstract

In Bayesian structure learning, we are interested in inferring a distribution over the directed acyclic graph (DAG) structure of Bayesian networks, from

of the Bayesian network, represented as a directed acyclic graph (DAG) and encoding the statistical dependencies between the variables of interest, is assumed to be known based on knowledge from domain experts. However, when this graph is unknown, we can learn the DAG structure of

Tristan Deleu, António Góis, Chris Chinenye Emezue, Mansi Rankawat, Simon Lacoste-Julien, Stefan Bauer, Yoshua Bengio, Bayesian Structure Learning with Generative Flow Networks ICLR 2022

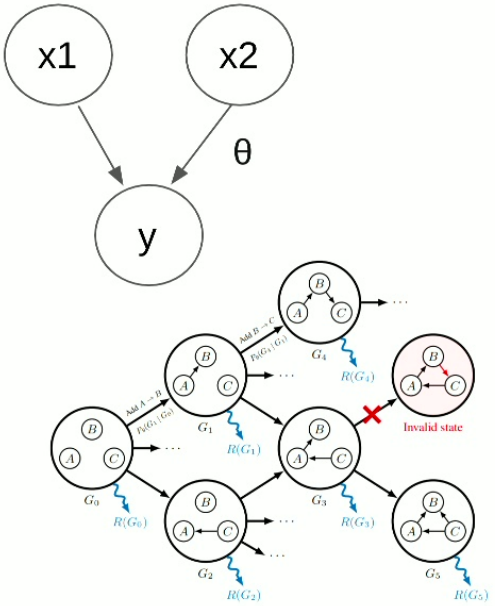
GFlowNets for causal structure learning (the first paper)

- Likelihood is given by the following

$$P(D|G) = \int P(D|G, \theta_G) P(\theta_G|G) d\theta_G$$

Gaussian likelihood

conjugate prior (gaussian)



$$R(G) = P(D \mid G) P(G)$$

We end up with a distribution over graphs but no information about the mechanisms. P(G) only.

GFlowNets for causal structure learning and inferring the mechanisms. (The second paper)

linear Gaussian assumption for the mechanism, and learned the parameters of the model and the graph. P(G, θ).

18

GFlowNets for causal structure learning and inferring the mechanisms. (The second paper)

VARIATIONAL BAYES DAG-GFLOWNET

Bayesian learning of Causal Structure and Mechanisms with GFlowNets and Variational Bayes

Mizu Nishikawa-Toomey* MIZU.NISHIKAWA-TOOMEY@MILA.QUEBEC
Mila, Université de Montréal

Tristan Deleu* DELEUTRI@MILA.QUEBEC
Mila, Université de Montréal

Jithendaraa Subramanian JITHENDARAA.SUBRAMANIAN@MILA.QUEBEC
Mila, McGill

Yoshua Bengio YOSHUA.BENGIO@MILA.QUEBEC
Mila, Université de Montréal

Laurent Charlin LCHARLIN@MILA.QUEBEC
Mila, HEC Montréal

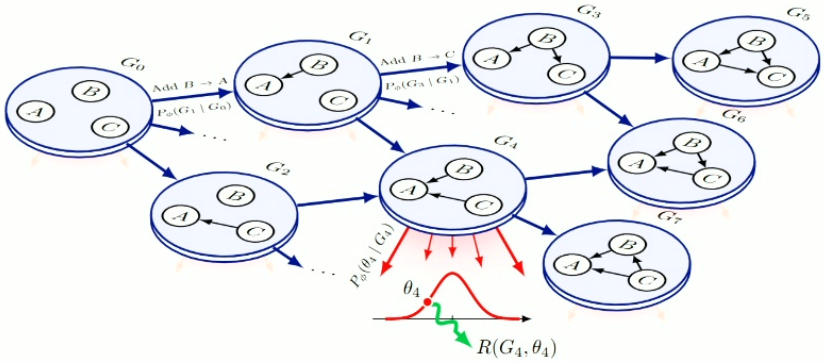
8 Apr 2023

GFlowNets for causal structure learning and inferring the mechanisms, with no assumptions on the mechanisms (the third paper).

Mila

Joint Bayesian Inference of Graphical Structure and Parameters with a Single Generative Flow Network

Tristan Deleu[1] Mizu Nishikawa-Toomey[1] Jithendaraa Subramanian[2]
Nikolay Malkin[1] Laurent Charlin[3] Yoshua Bengio[1,4]



023

19

Related work on causal structure learning

Bayesian:

- Dibs

  Lorch, Lars et al. "DiBS: Differentiable Bayesian Structure Learning." *ArXiv* abs/2105.11839 (2021): .

- BCD Nets

  Cundy, Chris et al. "BCD Nets: Scalable Variational Approaches for Bayesian Causal Discovery." *ArXiv* abs/2112.02761 (2021):

- DECI

  Geffner, Tomas et al. "Deep End-to-end Causal Inference." *ArXiv* abs/2202.02195 (2022):

Non-Bayesian

- DAGS with no tears

  Zheng, Xun et al. "DAGs with NO TEARS: Continuous Optimization for Structure Learning." *Neural Information Processing Systems* (2018).

- Gran-DAG

  Lachapelle, Sébastien et al. "Gradient-Based Neural DAG Learning." *ArXiv* abs/1906.02226 (2019):

- DAG-GNN

  Yu, Yue et al. "DAG-GNN: DAG Structure Learning with Graph Neural Networks." International Conference on Machine Learning (2019).

- AVICI:

  Lorch, Lars et al. "Amortized Inference for Causal Structure Learning." ArXiv abs/2205.12934 (2022): n

Empirical results on erdos-Renyi graphs

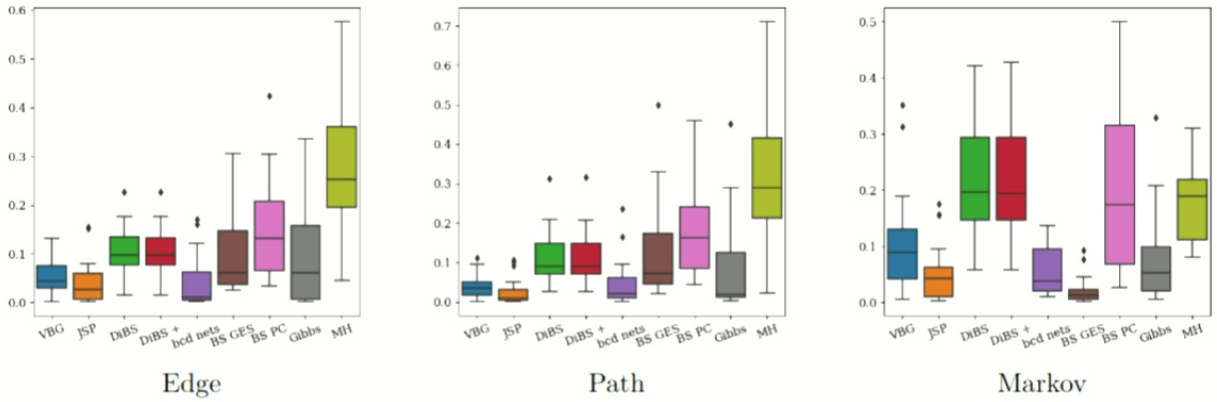


Figure 1: MSE of Edge, path and Markov features of the true posterior and the estimated posterior for 5 node Erdos-Renyi graphs (lower the better).

Nishikawa-Toomey, Mizu et al. "Bayesian learning of Causal Structure and Mechanisms with GFlowNets and Variational Bayes." *ArXiv* abs/2211.02763 (2022)

Comparison with the estimated and true posterior, GFlowNet based methods outperform others.

Challenges with GFlowNets

- Scalability.

Causal structure learning with GFlowNets has only been tested on graphs of size 50.

- Convergence.

We rely on the detailed balance loss to be minimised. Which at times has numerical issues.

- Inability to verify quality of uncertainty estimates for larger nodes.