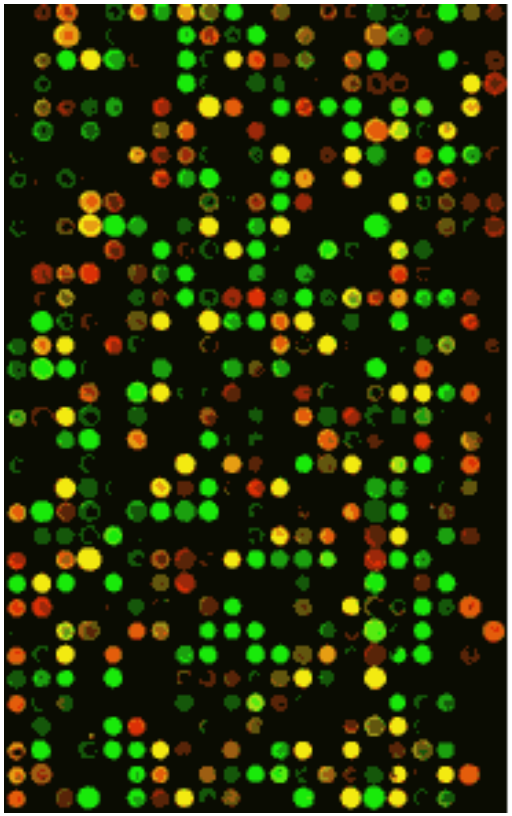
Evaluating the Effect of Perturbations in Reconstructing Network Topologies

Florian Markowetz and Rainer Spang
Max-Planck-Institute for Molecular Genetics
– Computational Molecular Biology –
Berlin, Germany
http://cmb.molgen.mpg.de/compdiag/

DSC 2003 Wien
Thursday, March 20

— Genetic networks —



- Microarrays provide a snapshot of gene expression in a cell. Genes are not expressed independently, they regulate each others activity.
- **Goal: Reconstruct the gene regulation network.**
- Clustering points to functional relationships, but fails to detect interactions between genes different from linear correlation.
- **Causality, not correlation!** Is the effect of a mutated gene on a target direct, or mediated by other genes? What is the nature of the interaction between genes (e.g. does gene A inhibit gene B)?

— Bayesian networks —

A **Bayesian Network** for $\mathbf{X} = \{X_1, \ldots, X_n\}$ consists of

- a **network structure** $\mathcal{G}$
  - directed acyclic graph (DAG),
  - nodes $\leftrightarrow$ variables,
  - lack of arc $\leftrightarrow$ conditional independence
- a set of **probability distributions** $\mathcal{P}$
  - locally: conditional distribution of a variable given its parents in the graph $\mathcal{G}$:

$$\mathcal{P} = \{ P(X_i \mid pa_i) \}$$

— Learning network structure —

1. **Constraint based:** construct graph by patterns of conditional independencies measured in the data (Pearl, SGS).

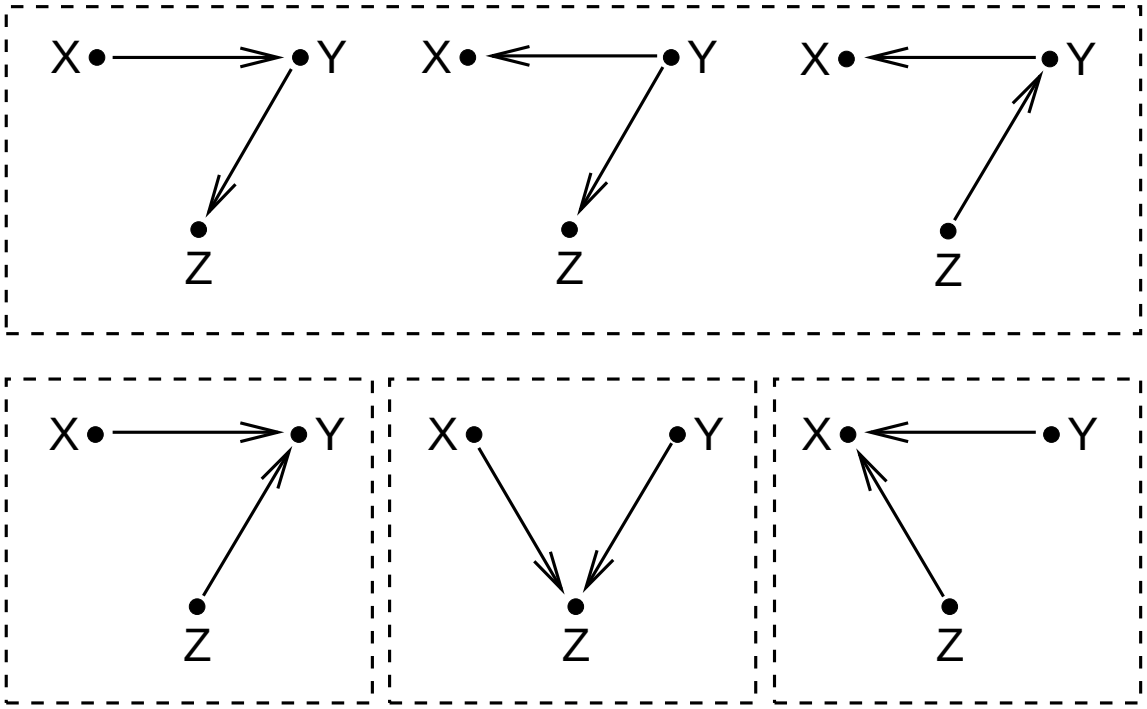
2. **Bayesian scoring:** use a scoring metric to evaluate the models and return the highest scoring model found (Heckerman).

$$P(\text{dag} \mid \text{data}) \propto P(\text{dag})\ P(\text{data} \mid \text{dag})$$

The number of graphs grows super-exponentially in the number of nodes. For more than 5 nodes a complete search is intractable (loophole: MCMC).

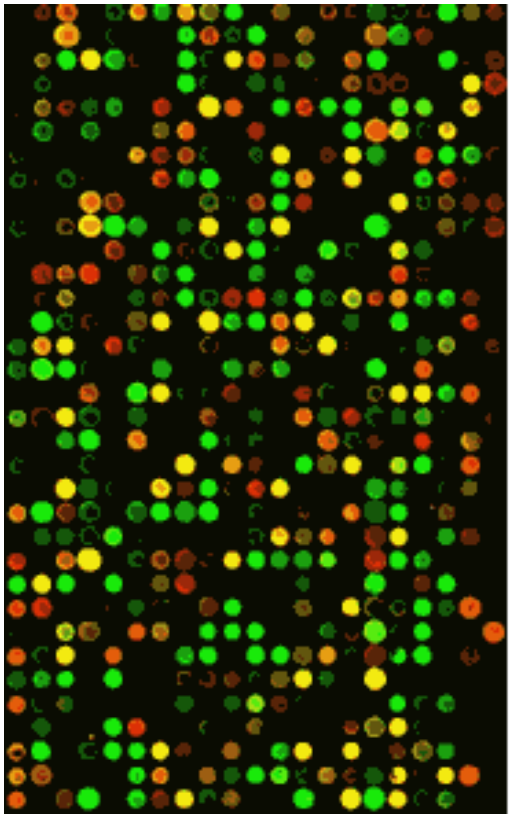
— Equivalence of bayesian networks —

Markov equivalence $\mathcal{G}_1 \overset{M}{\sim} \mathcal{G}_2$ if both structures represent the same set of independence assertions, i. e. if they are compatible with the same distribution P.



Even with infinitly many observations we cannot decide between the DAGs in the same equivalence class.

— Observation and Intervention —



What effect does this result have on the reconstruction of genetic networks by BN?

- Arrows in the BN do not necessarily represent causal influence! From observations alone we can only learn whole equivalence classes, in general not a single DAG.

- But biologists not only observe, they also **intervene, perturb, disrupt** the gene network e. g. by knock-out experiments.

- **How much do we gain by using perturbation data in structure learning?**

— Objectives —

Evaluating the effect of interventional data on learning network structure.

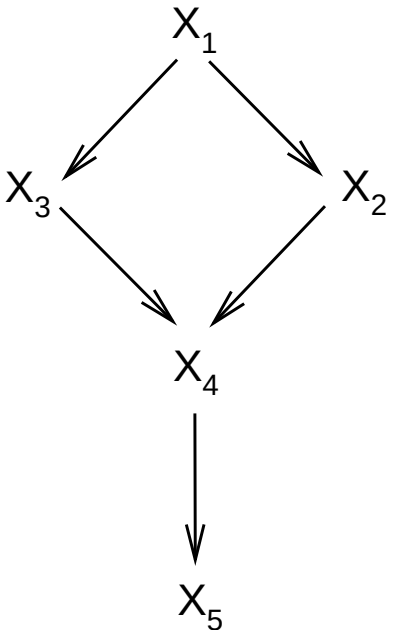
Starting point: small network of 5 nodes with 3 states each. Sample data with and without interventions (100, 50, 25 observations). Then reconstruct the original topology by bayesian scoring (exhaustive search).

1. **Score distribution:** is the DAG with maximal score singled out sharply, or are there other DAGs with almost the same high score?

2. **Sample size:** How many data are needed to correctly identify the underlying structure?

3. **Robustness:** How does the learning accuracy vary with changes in the conditional probability distributions?

— The κ-network —



A small simulation network: 5 nodes, 3 states each.

Topology of the sprinkler network.

Conditional probability distributions are multinomials depending on a parameter κ by the scheme:

$$\kappa \cdot \text{signal} + (1-\kappa) \cdot \text{noise}$$

$$T_{(pa=1)} = \kappa \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \frac{1-\kappa}{3} \cdot (\text{ ones })_{3\times 3}$$

— Experimental setup —

1. We selected κ from 0 to 0.9 in steps of 0.1.

   The value $\kappa = 1$ is omitted because the lack of random effects results in learning a completely connected graph.

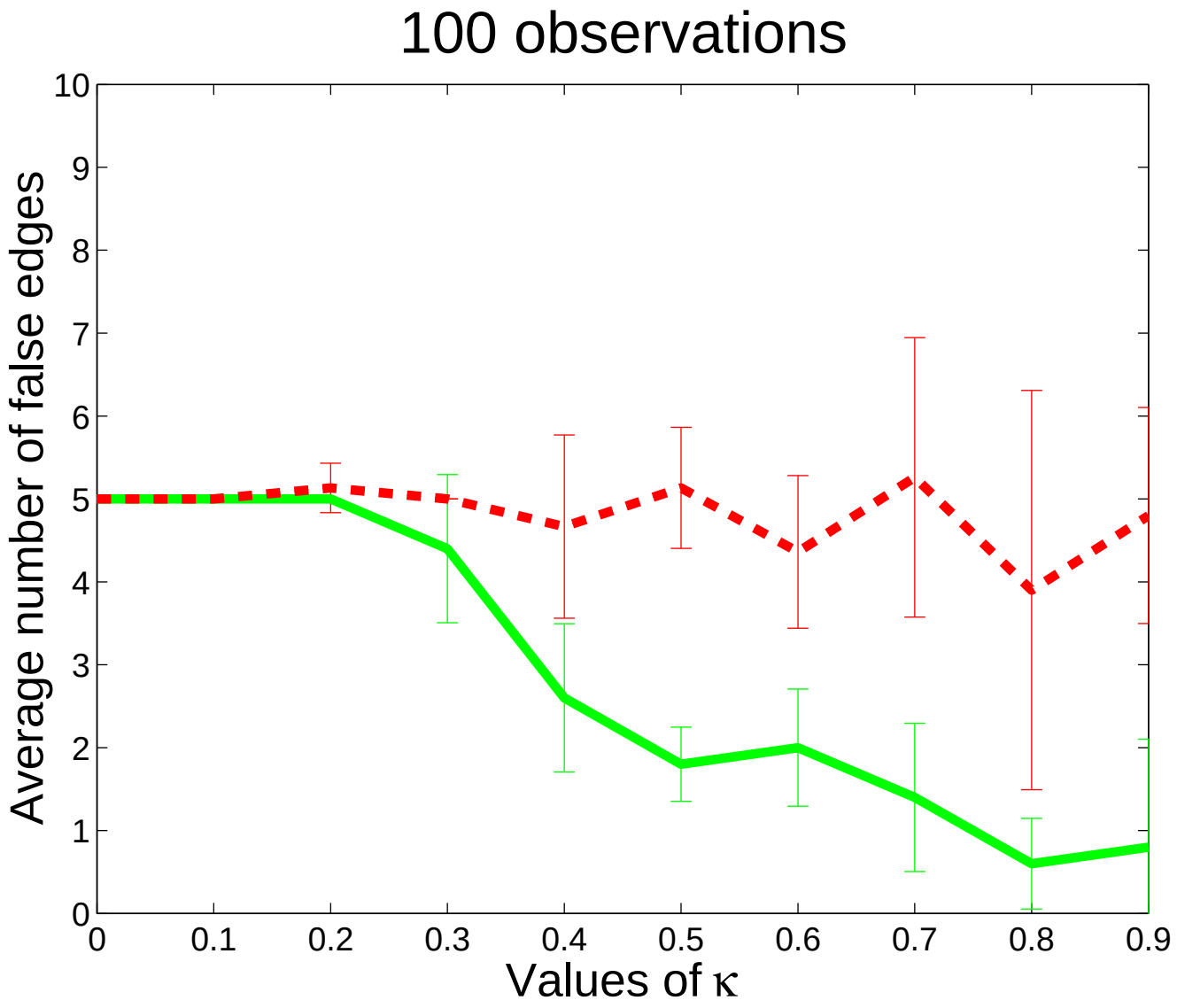
2. We sampled data from the κ-network and searched for DAGs with high bayesian score by exhaustive search over all 29281 model structures.

3. Average over all DAGs with highest score.

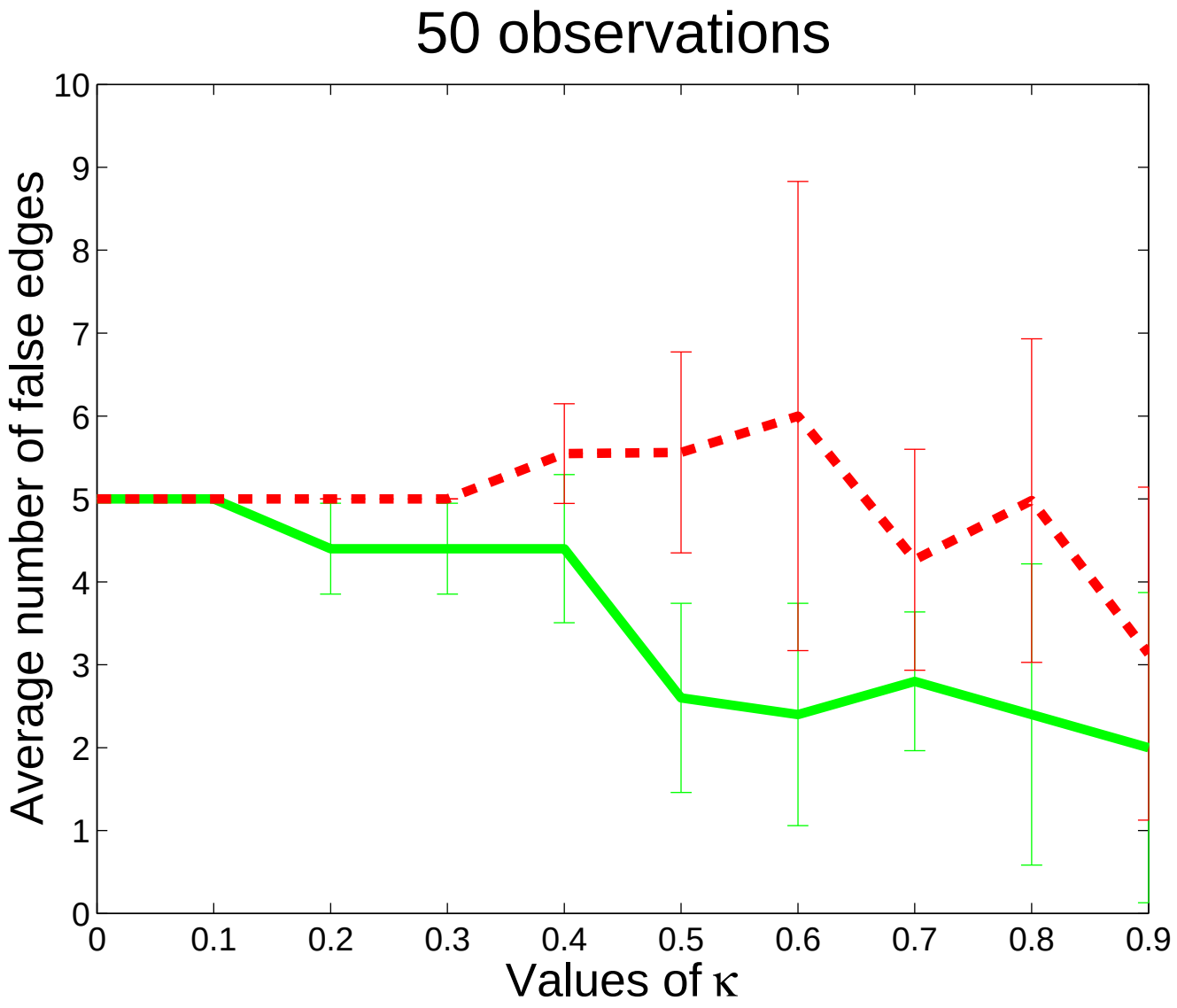
4. Distance to the true topology $=$ number of falsely predicted edges.

5. We repeated the whole process 5 times and took the average of the number of falsely predicted edges.

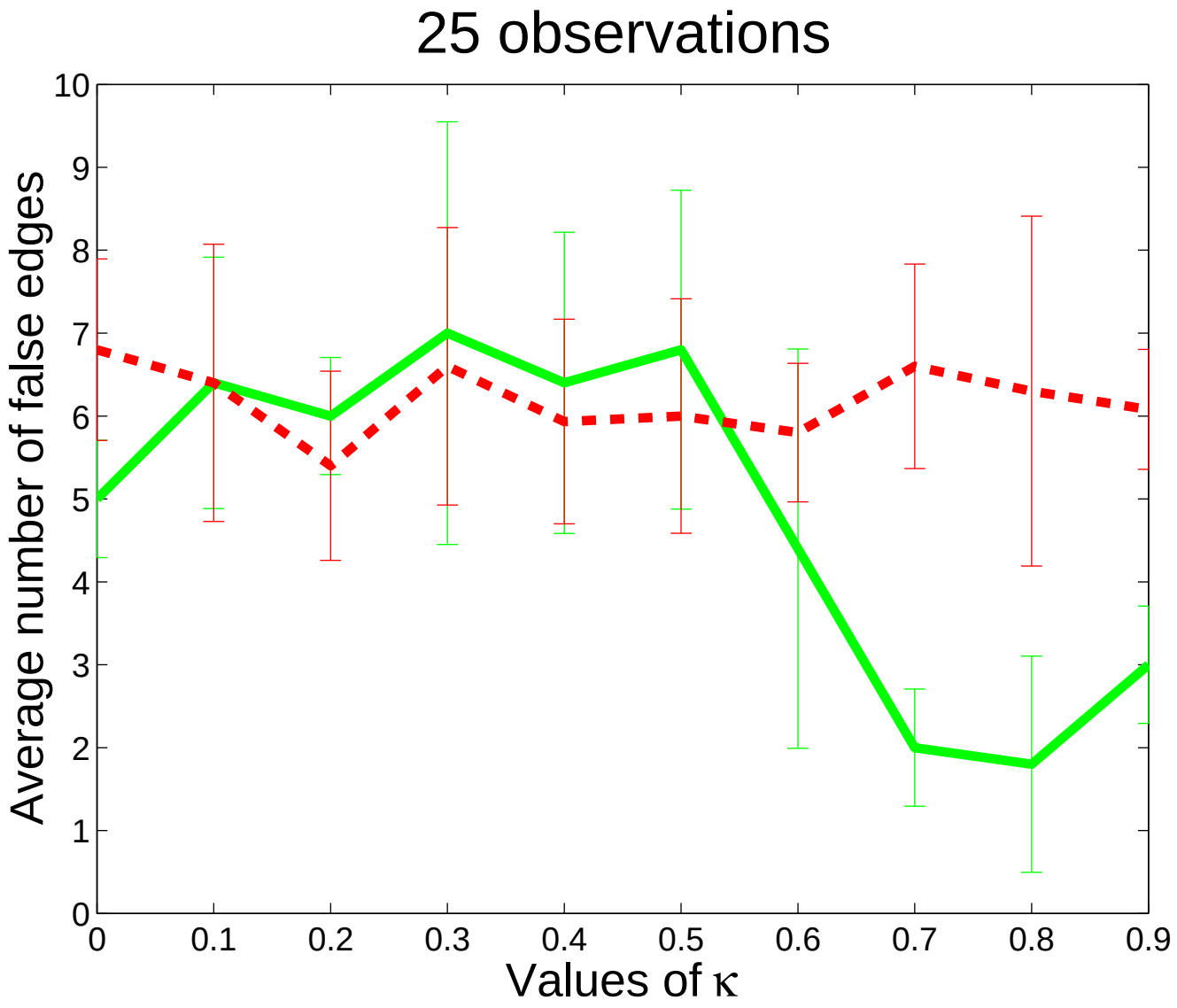
— **Without interventions** – **With interventions** —



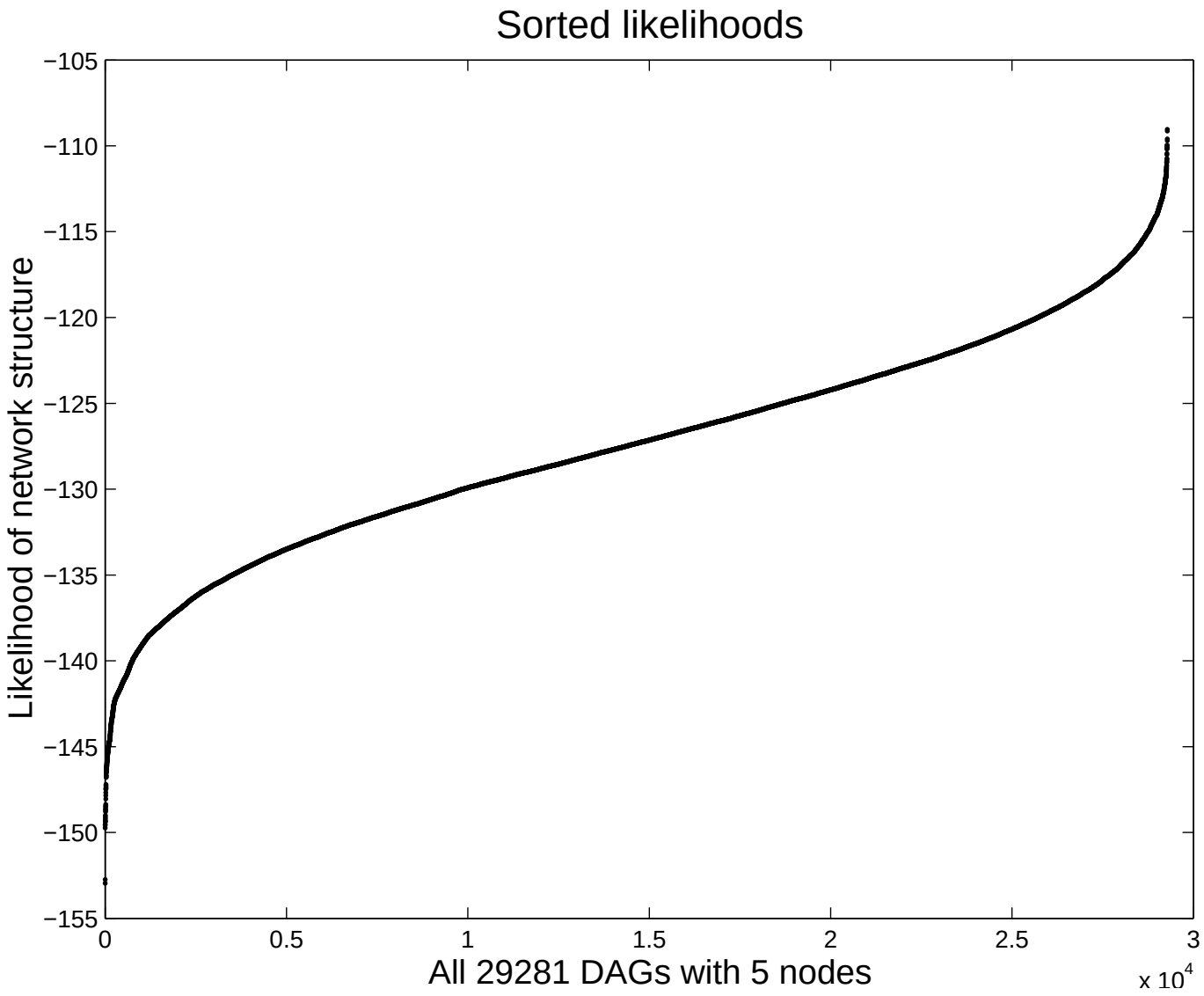
— **Without interventions** – **With interventions** —



— Without interventions – With interventions —



— Score distribution of all DAGs with 5 nodes —



— Results —

- Data from perturbation experiments increases the accuracy of structure learning.
- The clearer the signal, the greater the difference between learning with and without interventions.
- Still, large datasets are needed to identify even small networks.

Aim only at small networks and
use data from perturbation experiments

— Further Research —

- Do we gain more information by perturbing more than one node?
- Experimental design: choose the next intervention such that the most additional information is gained.