

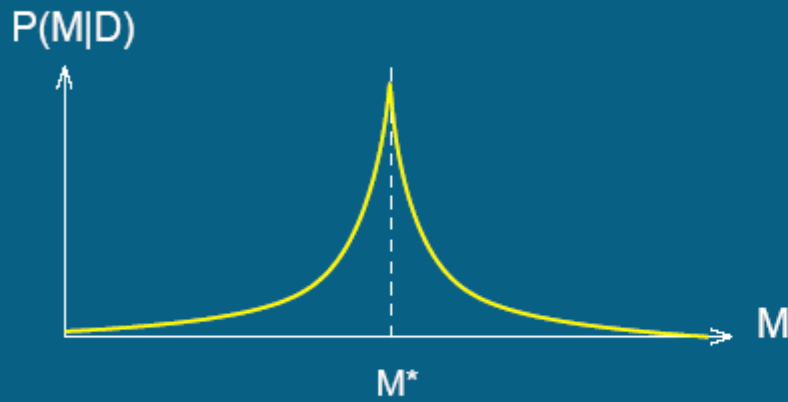
Learning Bayesian networks with improved MCMC schemes

Dirk Husmeier

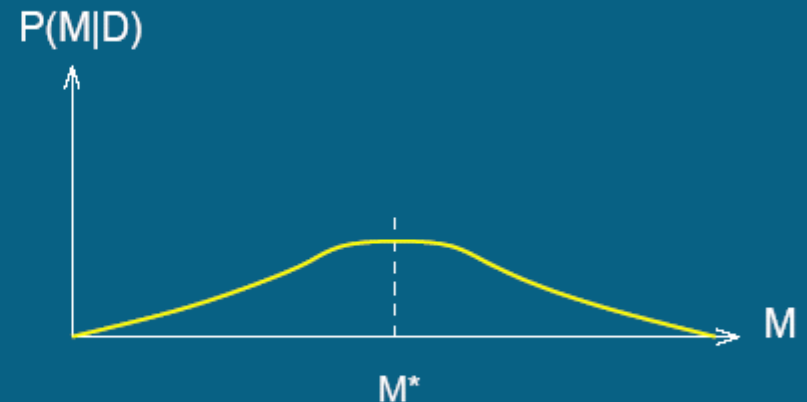
Biomathematics & Statistics Scotland

Learning Bayesian networks

Data are sparse \rightarrow Intrinsic uncertainty of inference



Large data set D:
Best network structure M^* well defined



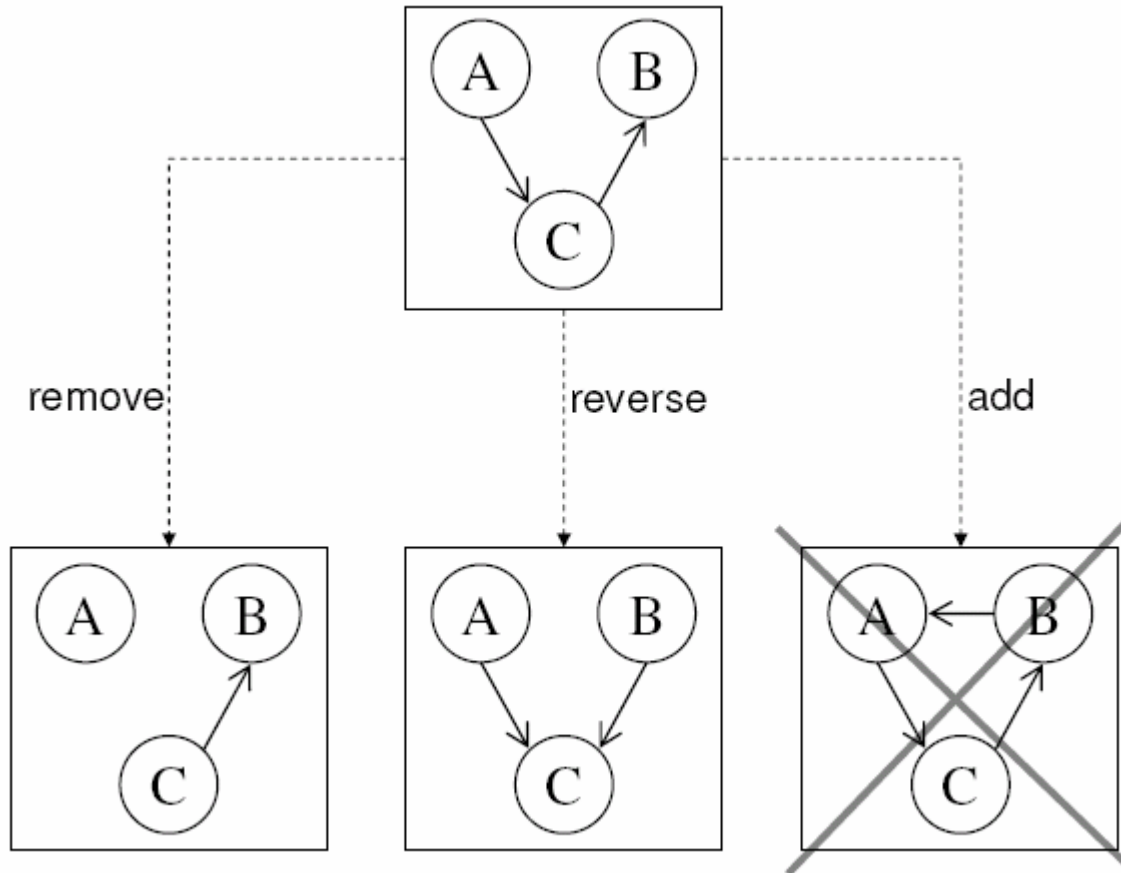
Small data set D:
Intrinsic uncertainty about M^*

$$P(M|D) = P(D|M) P(M) / Z$$

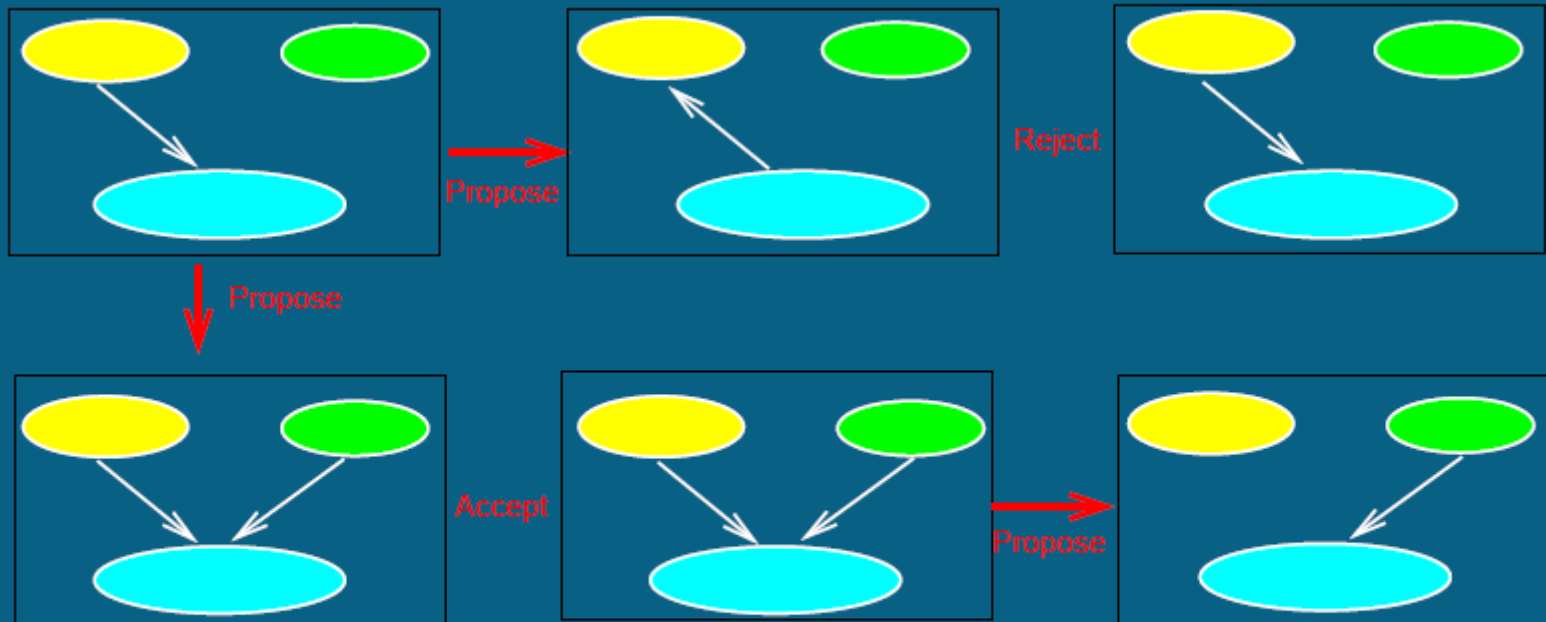
M: Network structure. D: Data

MCMC in structure space

Madigan & York (1995), Guidici & Castello (2003)



Markov chain Monte Carlo (MCMC)



Acceptance probability: $\min \left\{ 1, \frac{P(D|M_{new})}{P(D|M_{old})} \times \frac{P(M_{new})}{P(M_{old})} \times \frac{Q(M_{old}|M_{new})}{Q(M_{new}|M_{old})} \right\}$

Alternative paradigm: order MCMC

Being Bayesian About Network Structure

A Bayesian Approach to Structure Discovery in Bayesian Networks

Nir Friedman (nir@cs.huji.ac.il)

School of Computer Science & Engineering

Hebrew University

Jerusalem, 91904, Israel

Daphne Koller (koller@cs.stanford.edu)

Computer Science Department

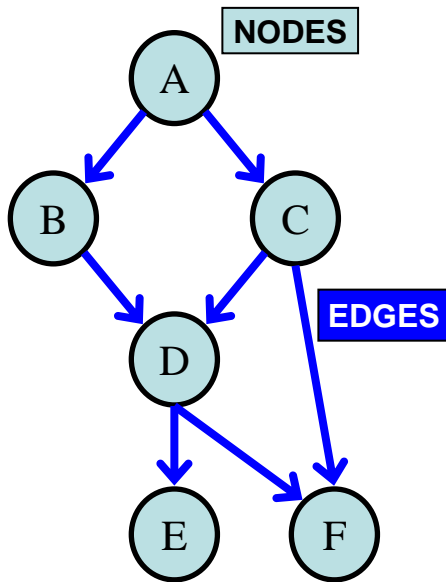
Stanford University

Stanford, CA 94305-9010

Machine Learning, 2004

Exploiting the modularity of Bayesian networks

$$P(G | D) \propto P(D | G)P(G) = \prod_i \text{score}(X_i, \text{Pa}_G(X_i) | D)$$

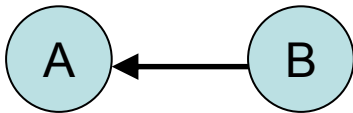
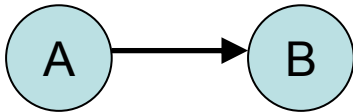
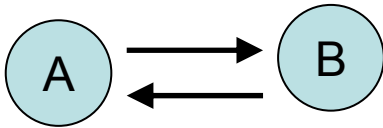


$$P(\text{Pa}_G(X_i) = \mathbf{U} | D, \dots) = \frac{\text{score}(X_i, \mathbf{U} | D)}{\sum_{\mathbf{U}' \in \mathcal{U}_{i,\dots}} \text{score}(X_i, \mathbf{U}' | D)}$$

$$P(A, B, C, D, E, F)$$

$$= P(A) \cdot P(B | A) \cdot P(C | A) \cdot P(D | B, C) \cdot P(E | D) \cdot P(F | C, D)$$

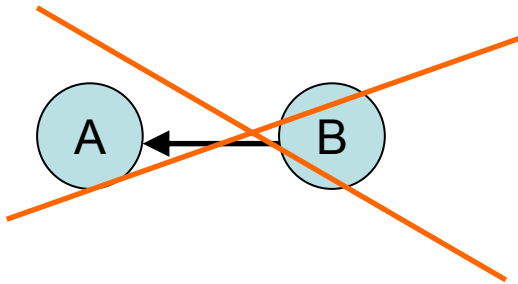
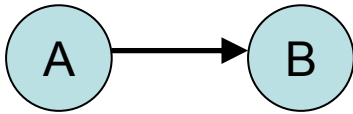
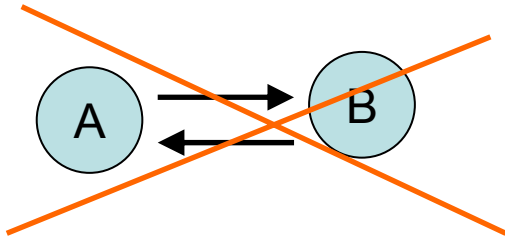
Possible structures



Two nodes:



Possible structures



Order constraint



Parents have to be “upstream” in the order.

Alternative paradigm: order MCMC

$$P(G | D) \propto P(D | G)P(G) = \prod_i \text{score}(X_i, \text{Pa}_G(X_i) | D)$$

$$P(\text{Pa}_G(X_i) = \mathbf{U} | D, \prec) = \frac{\text{score}(X_i, \mathbf{U} | D)}{\sum_{\mathbf{U}' \in \mathcal{U}_{i, \prec}} \text{score}(X_i, \mathbf{U}' | D)}$$

$$\mathcal{U}_{i, \prec} = \{\mathbf{U} : \mathbf{U} \prec X_i, |\mathbf{U}| \leq k\}.$$

where $\mathbf{U} \prec X_i$ is defined to hold when all nodes in \mathbf{U} precede X_i in \prec

$$\prec \quad \mapsto \quad (i_1 \dots i_j \dots i_k \dots i_n)$$

We first consider the problem of computing the probability of the data given the order:

$$P(D | \prec) = \sum_{G \in \mathcal{G}_k} P(G | \prec) P(D | G) \quad (7)$$

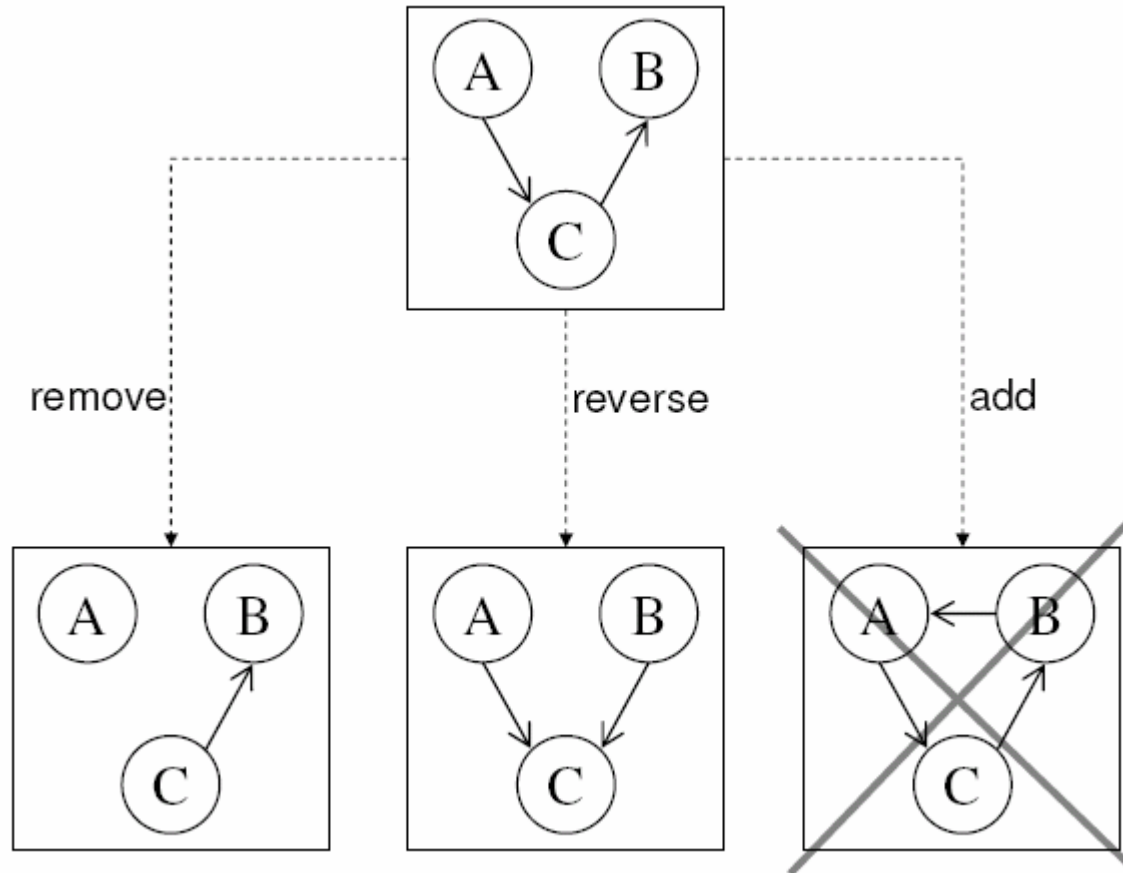
Given our constraint on the size of the family, the possible parent sets for the node X_i is

$$\mathcal{U}_{i, \prec} = \{\mathbf{U} : \mathbf{U} \prec X_i, |\mathbf{U}| \leq k\}.$$

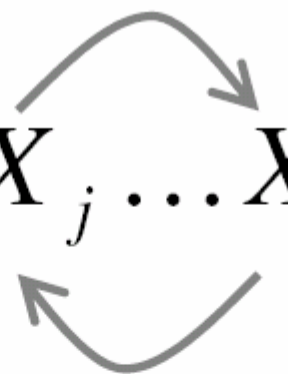
where $\mathbf{U} \prec X_i$ is defined to hold when all nodes in \mathbf{U} precede X_i in \prec . Let $\mathcal{G}_{k, \prec}$ be the set of structures in \mathcal{G}_k consistent with \prec . Using Eq. (5), we have that

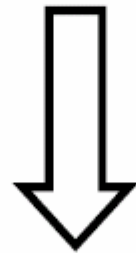
$$\begin{aligned} P(D | \prec) &= \sum_{G \in \mathcal{G}_{k, \prec}} \prod_i \text{score}(X_i, \text{Pa}_G(X_i) | D) \\ &= \prod_i \sum_{\mathbf{U} \in \mathcal{U}_{i, \prec}} \text{score}(X_i, \mathbf{U} | D). \end{aligned} \quad (8)$$

Instead of MCMC in structure space



MCMC in order space

$$\mathcal{L}_{old} = \{X_1 \dots X_j \dots X_k \dots X_N\}$$




$$\mathcal{L}_{new} = \{X_1 \dots X_k \dots X_j \dots X_N\}$$

MCMC method:

$$(i_1 \dots i_j \dots i_k \dots i_n) \mapsto (i_1 \dots i_k \dots i_j \dots i_n)$$

It remains only to discuss the construction of the Markov chain. We use a standard Metropolis algorithm (Metropolis et al., 1953). We need to guarantee two things:

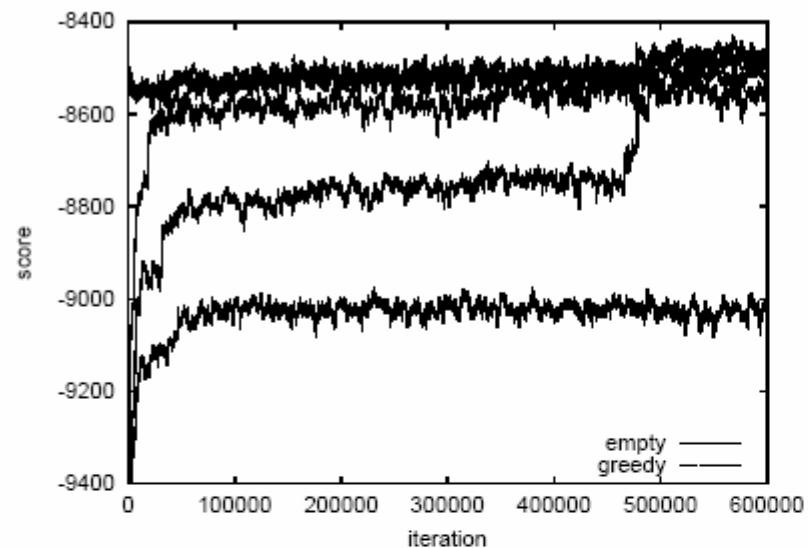
- that the chain is *reversible*, i.e., that $P(\prec \mapsto \prec') = P(\prec' \mapsto \prec)$;
- that the stationary distribution of the chain is the desired posterior distribution $P(\prec | D)$.

We accomplish this goal using a standard Metropolis sampling. For each order \prec , we define a *proposal probability* $q(\prec' | \prec)$, which defines the probability that the algorithm will “propose” a move from \prec to \prec' . The algorithm then *accepts* this move with probability

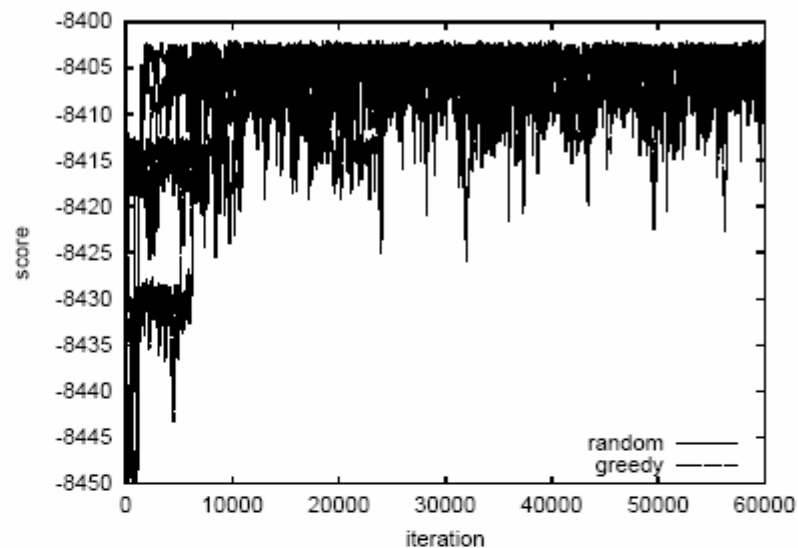
$$\min \left[1, \frac{P(\prec' | D)q(\prec | \prec')}{P(\prec | D)q(\prec' | \prec)} \right]$$

Structure

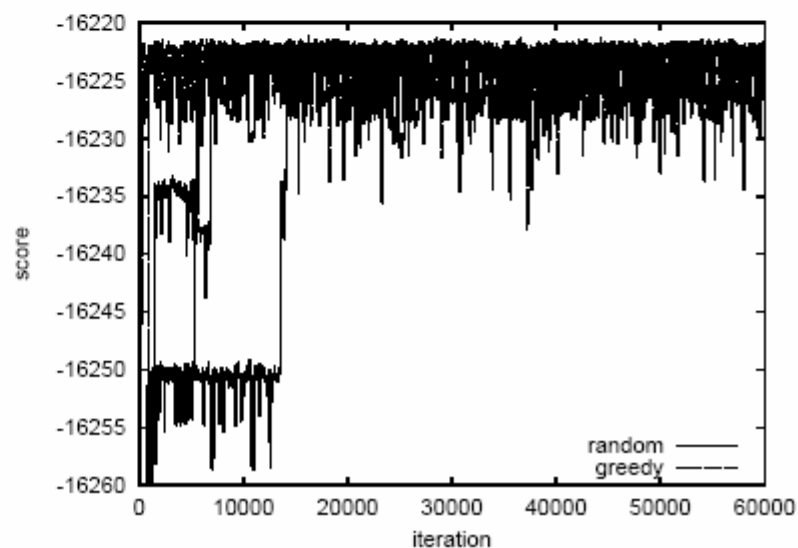
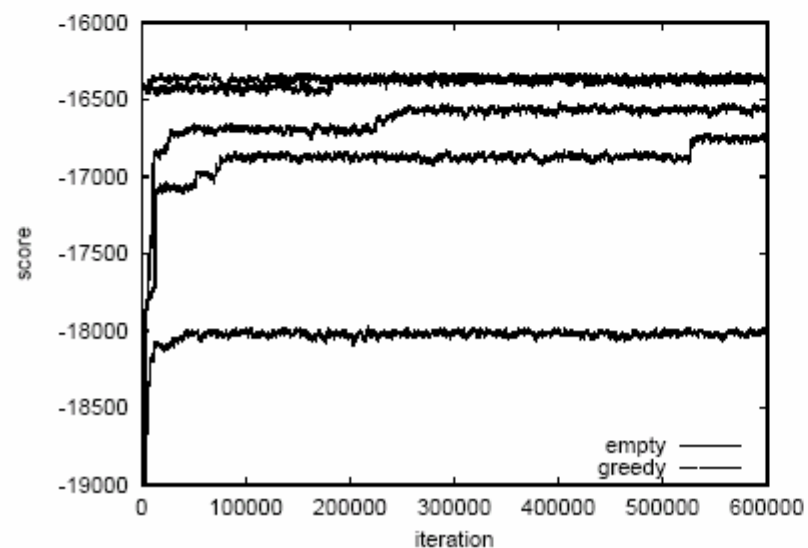
500 Instances



Order



1000 Instances

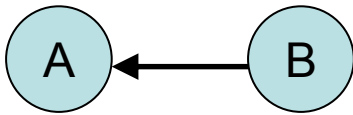
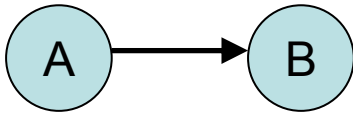


Problem:

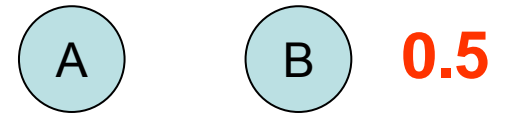
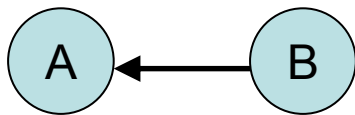
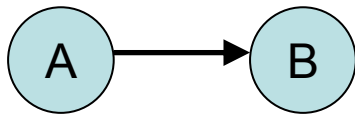
Distortion of the prior distribution

We introduce a uniform prior over orders \prec , and define $P(G | \prec)$ to be of the same nature as the priors we used in the previous section. It is important to note that the resulting prior over structures has a different form than our original prior over structures. For example, if we define $P(G | \prec)$ to be uniform, we have that $P(G)$ is not uniform: graphs that are consistent with more orders are more likely. For example, a Naive Bayes graph is consistent with $(n - 1)!$ orders, whereas any chain-structured graph is consistent with only one. As one consequence, our induced structure distribution is not *hypothesis equivalent* (Heckerman et al., 1995), in that different network structures that are in the same equivalence class often have different priors. For example, the chain $X \rightarrow Y \rightarrow Z$ is associated with a unique order, whereas the equivalent structure $X \leftarrow Y \rightarrow Z$ is associated with two orders, and is therefore twice as likely a priori.

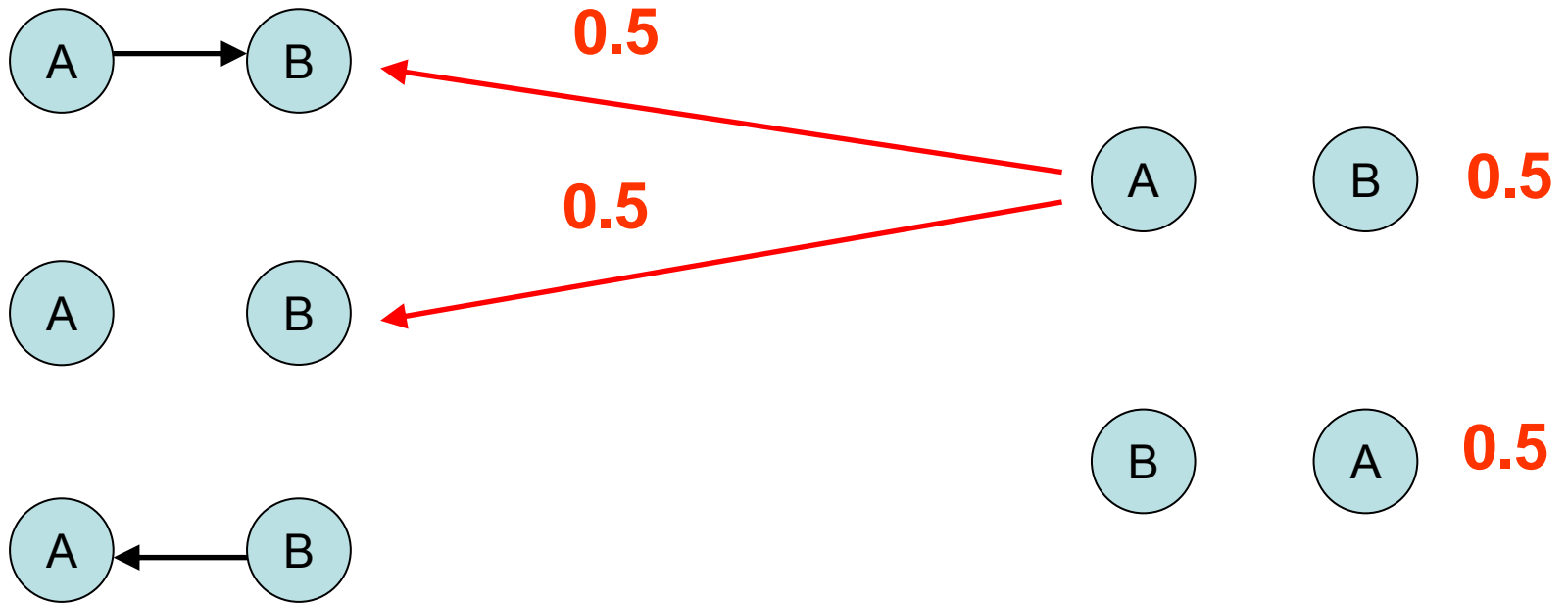
$$P(\mathcal{M}) = \sum_{\prec} P(\mathcal{M} | \prec) \cdot P(\prec)$$



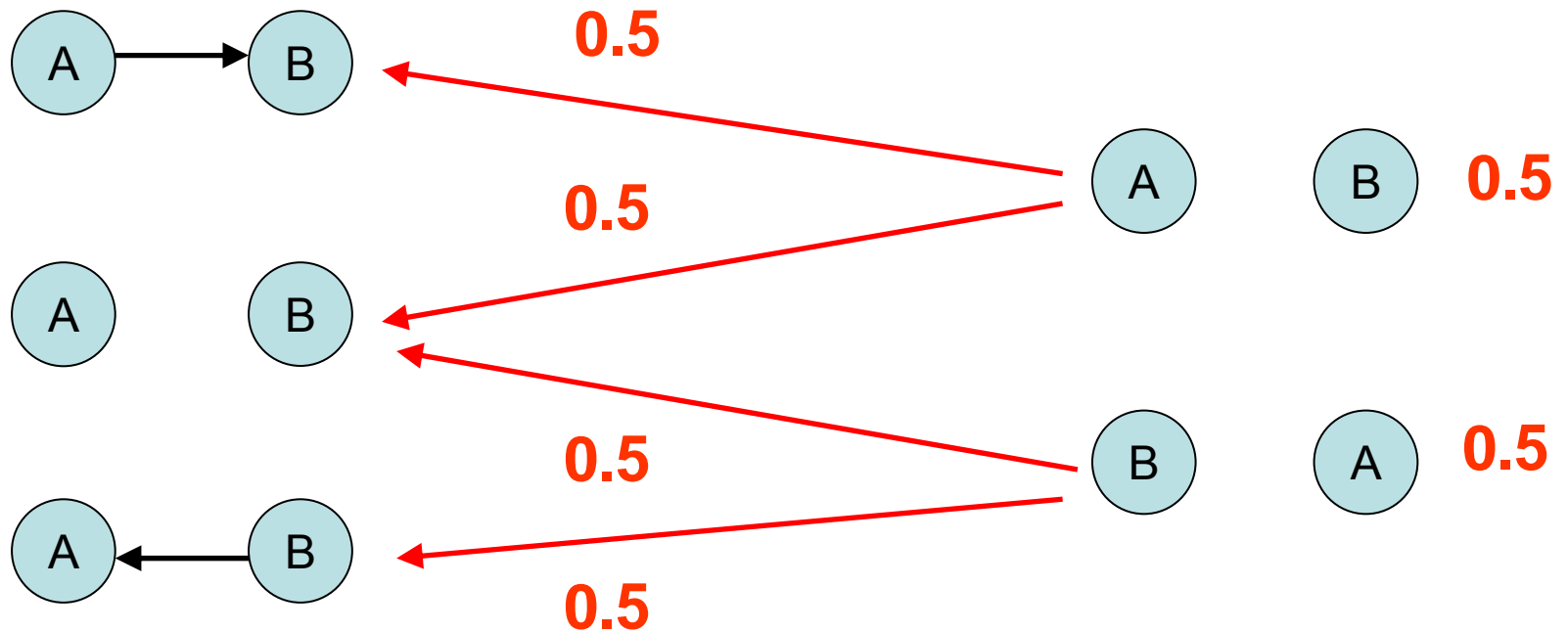
$$P(\mathcal{M}) = \sum_{\prec} P(\mathcal{M} | \prec) \cdot P(\prec)$$



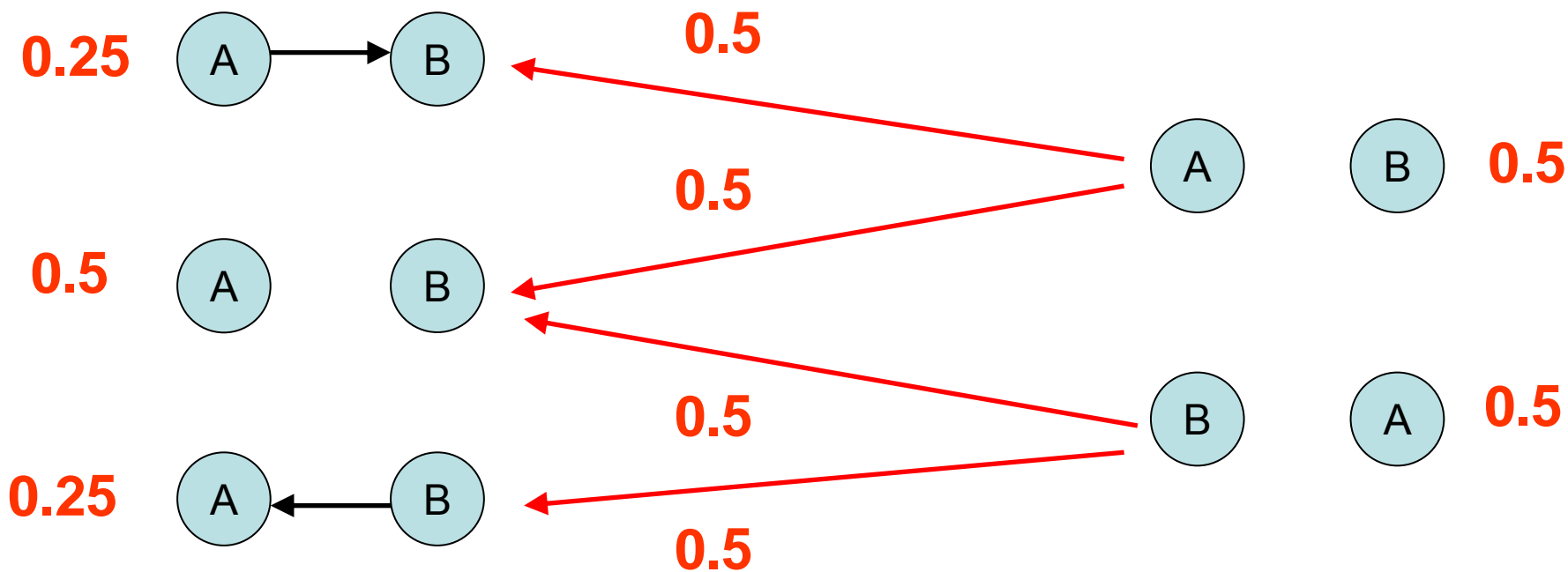
$$P(\mathcal{M}) = \sum_{\prec} P(\mathcal{M} | \prec) \cdot P(\prec)$$



$$P(\mathcal{M}) = \sum_{\prec} P(\mathcal{M} | \prec) \cdot P(\prec)$$



$$P(\mathcal{M}) = \sum_{\prec} P(\mathcal{M} | \prec) \cdot P(\prec)$$



Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move

Marco Grzegorzcyk · Dirk Husmeier

Received: 27 March 2007 / Revised: 11 January 2008 / Accepted: 28 March 2008
Springer Science+Business Media, LLC 2008

PROOF

Proposed new paradigm

- MCMC in **structure space** rather than order space.
- Design **new proposal moves** that achieve faster mixing and convergence.

Idea

Propose new parents from the distribution:

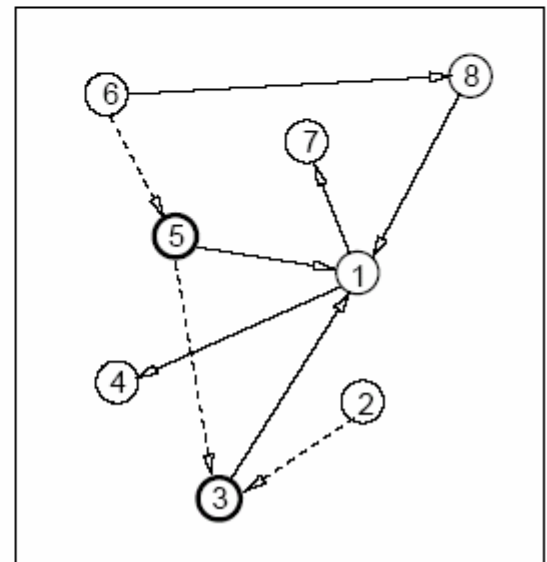
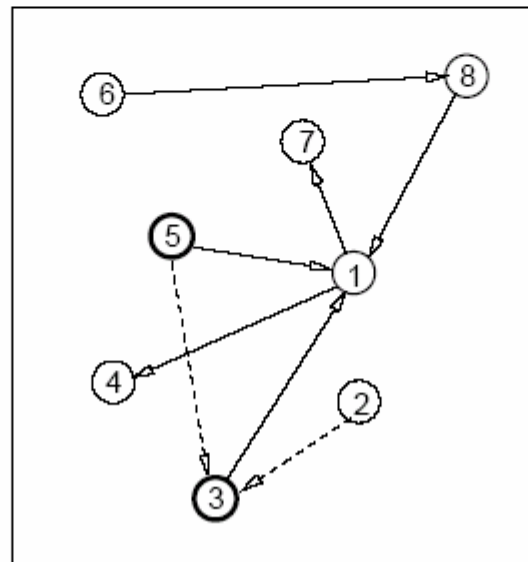
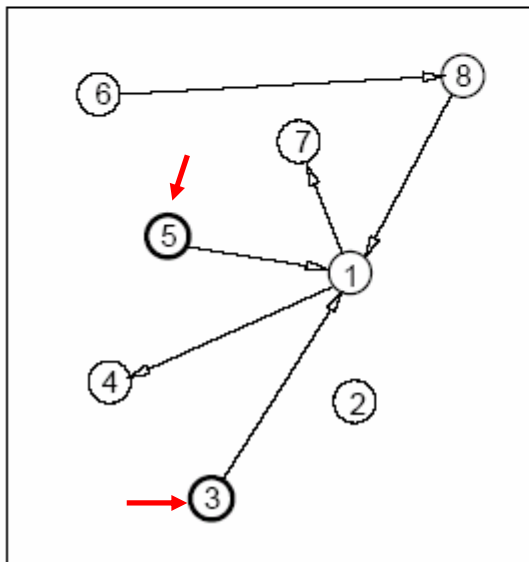
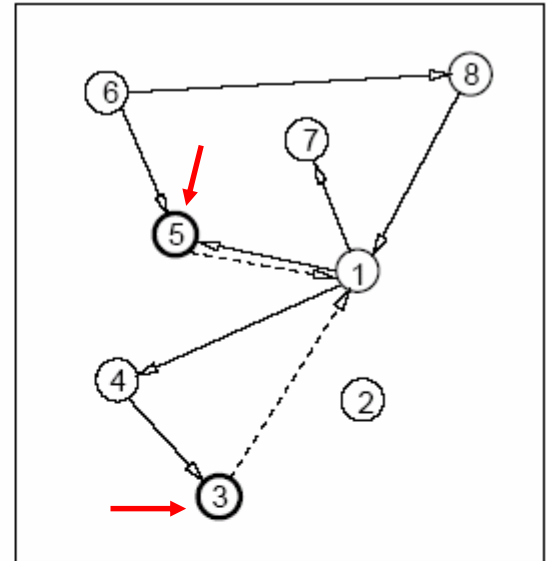
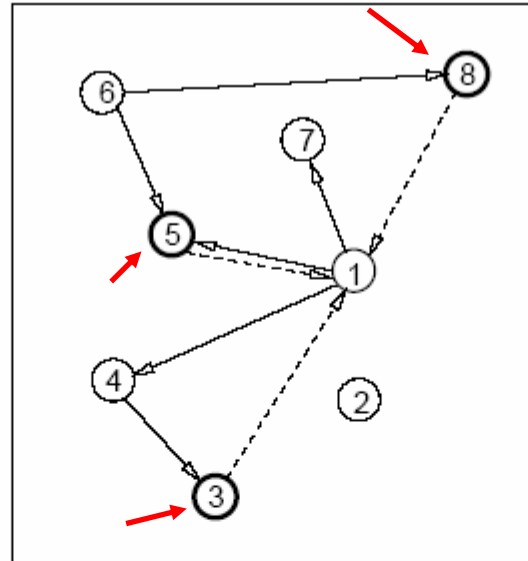
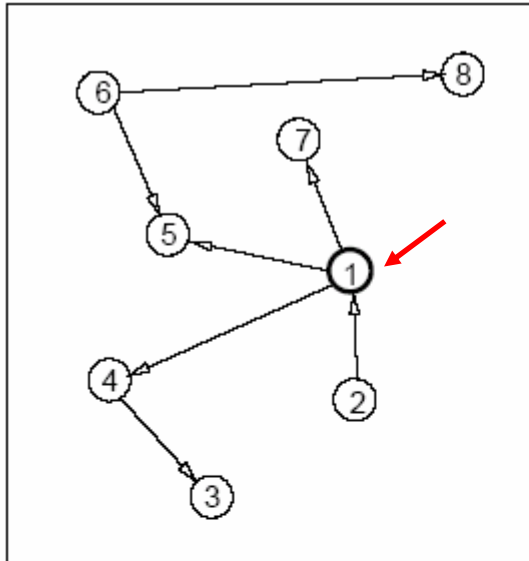
$$P(\text{Pa}_G(X_i) = \mathbf{U} \mid D, \times) = \frac{\text{score}(X_i, \mathbf{U} \mid D)}{\sum_{\mathbf{U}' \in \mathcal{U}_{i, \times}} \text{score}(X_i, \mathbf{U}' \mid D)}$$

- **Identify** those new **parents** that are involved in the formation of **directed cycles**.
- **Orphan** them, and **sample new parents** for them subject to the **acyclicity constraint**.

1) Select a node

2) Sample new parents

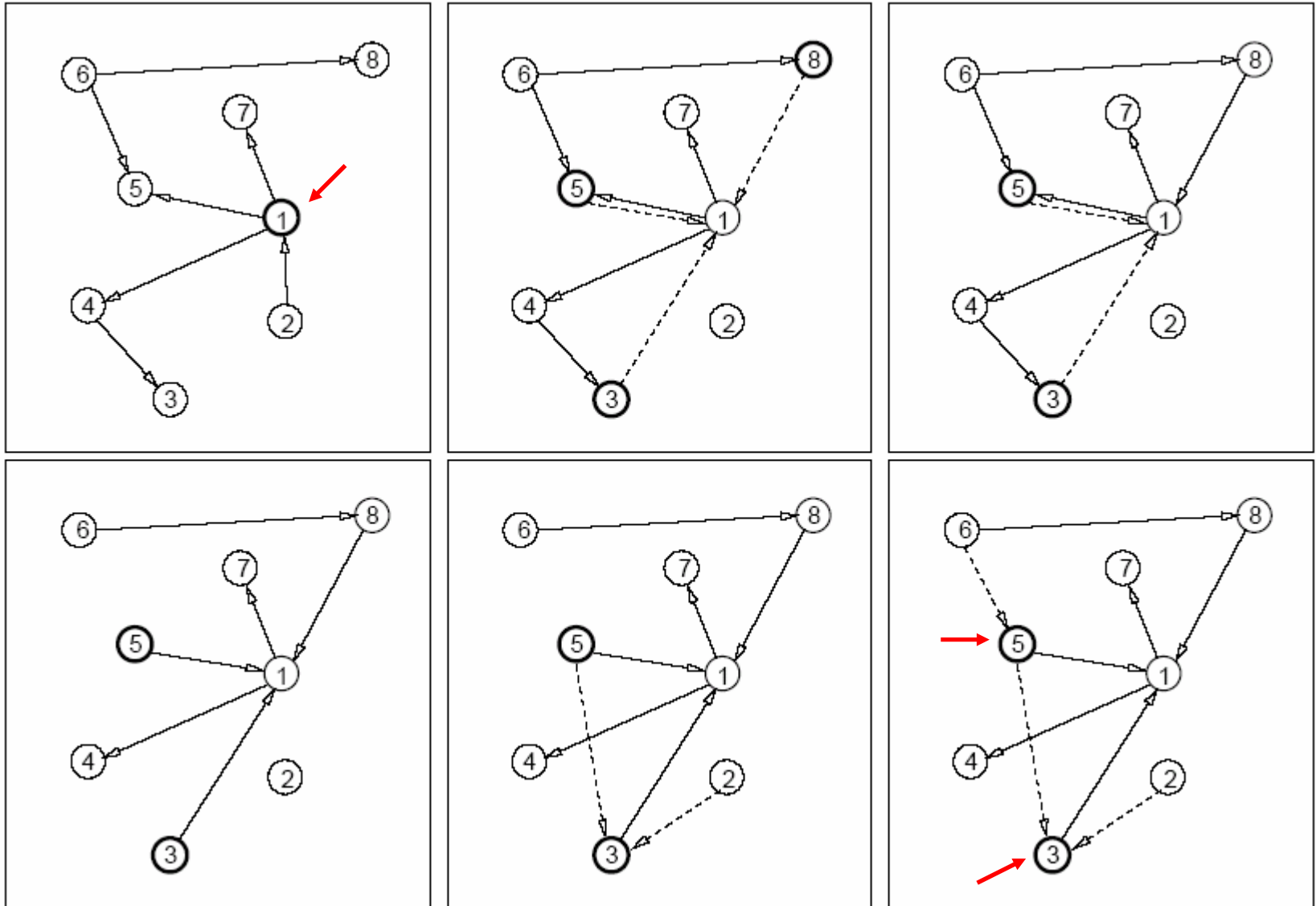
3) Find directed cycles



4) Orphan "loopy" parents

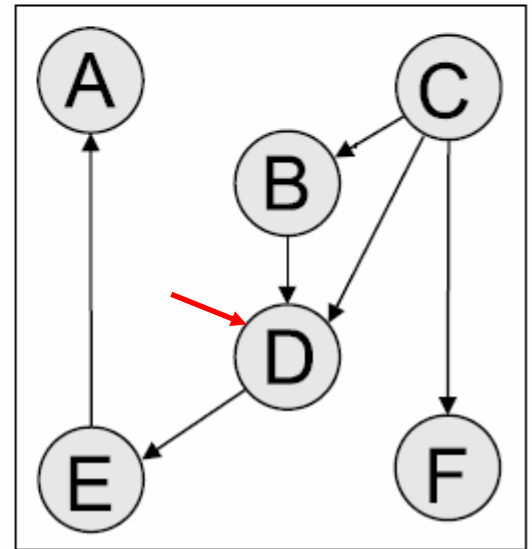
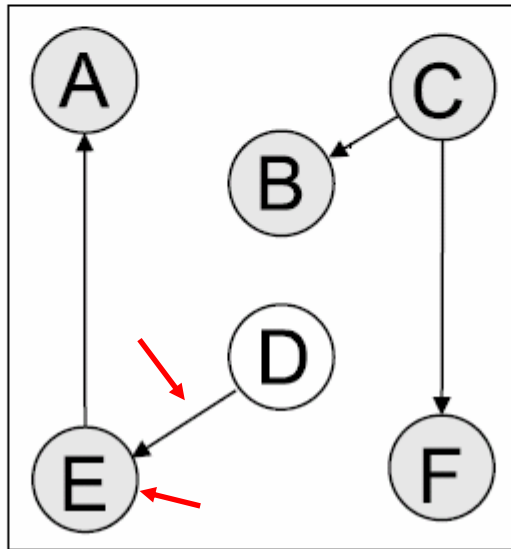
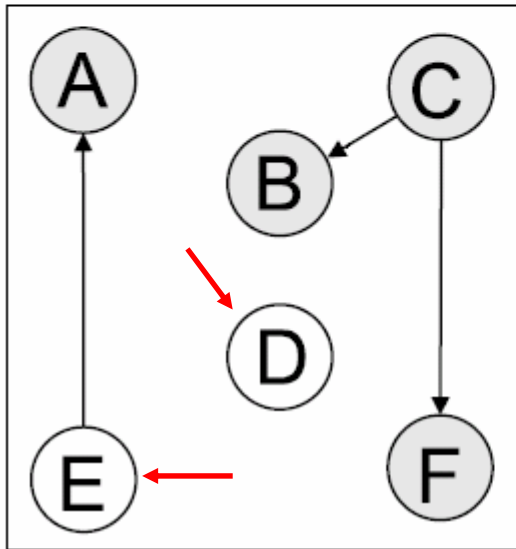
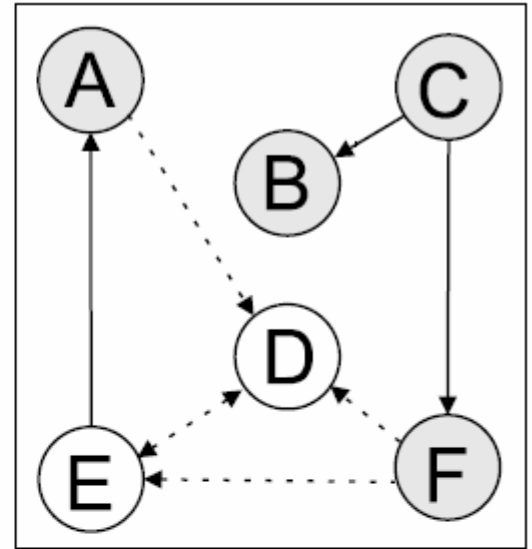
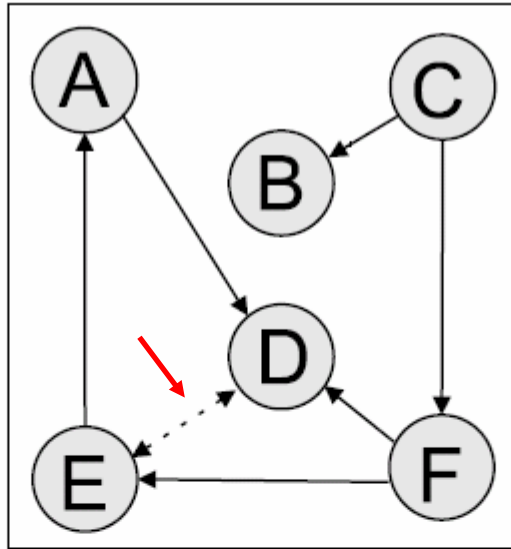
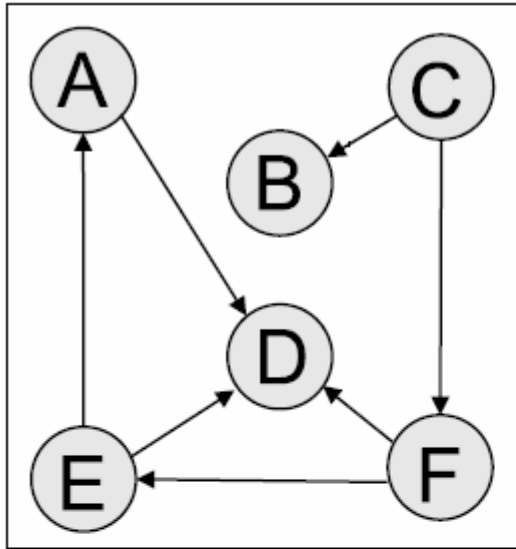
5) Sample new parents for these parents

Path via illegal structure



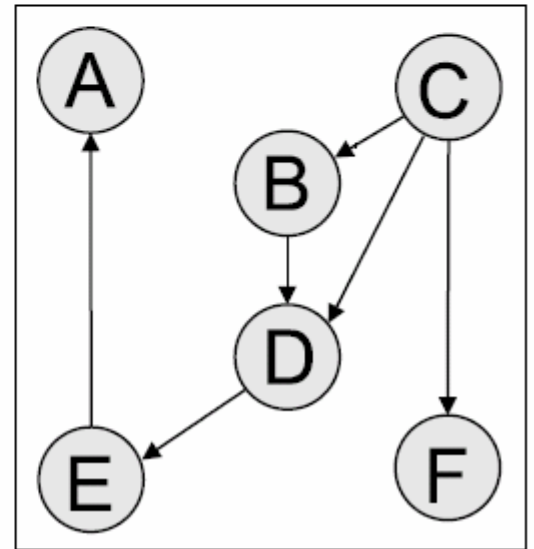
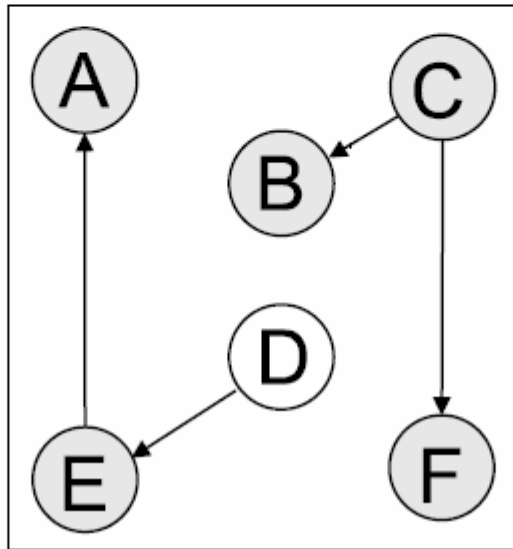
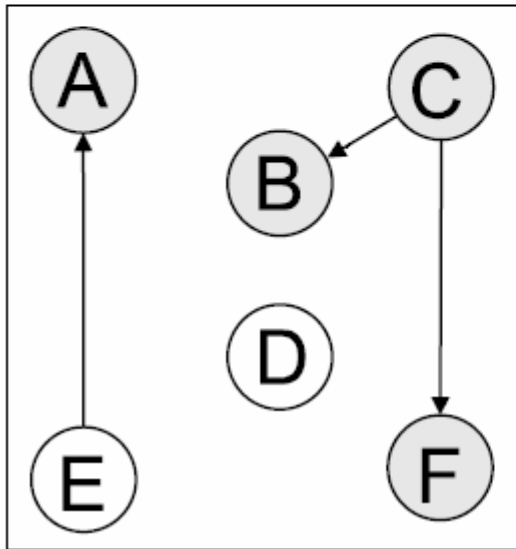
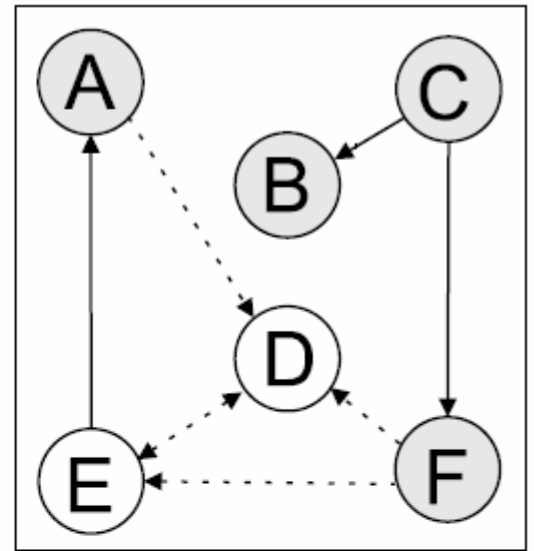
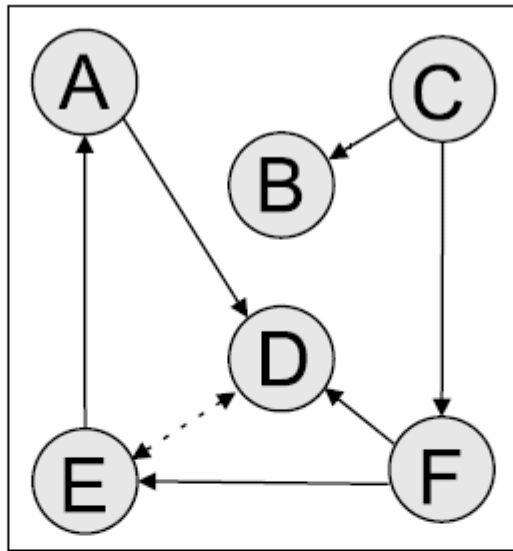
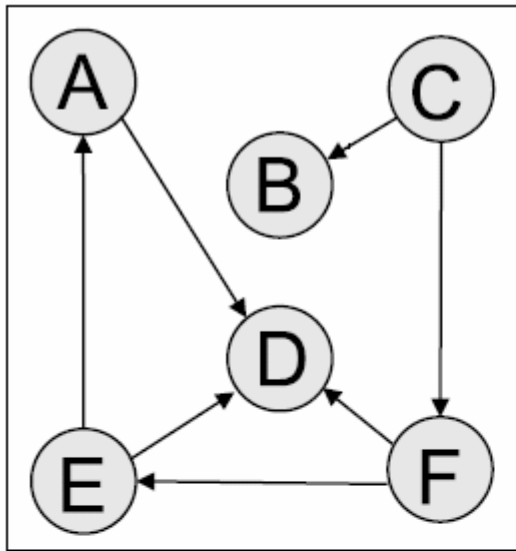
Problem: This move is not reversible

1) Select an edge



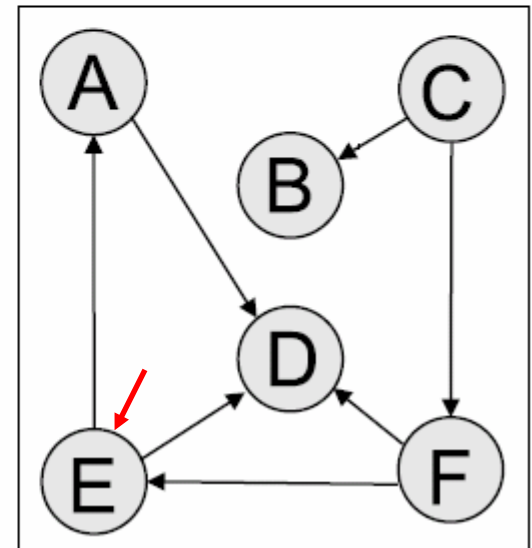
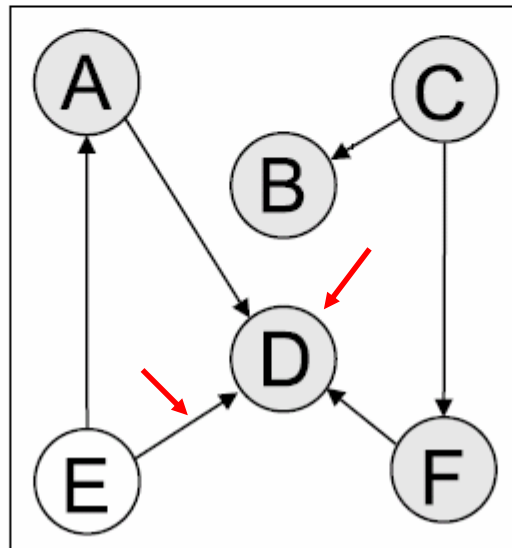
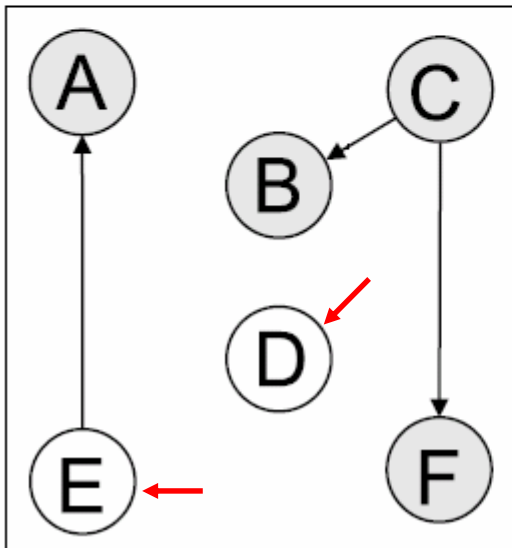
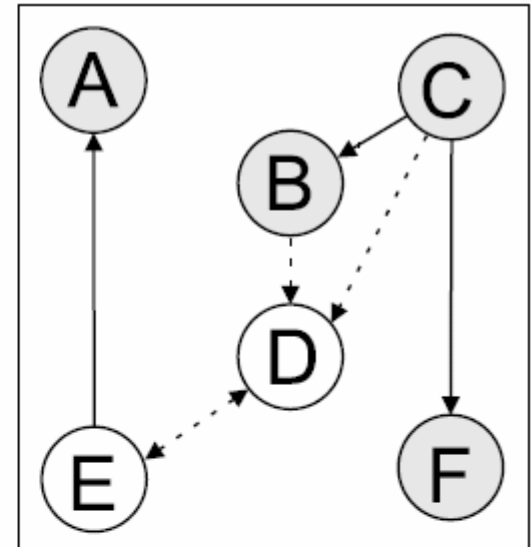
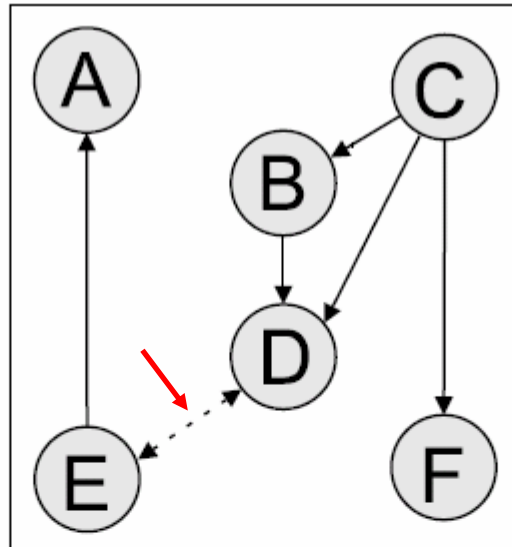
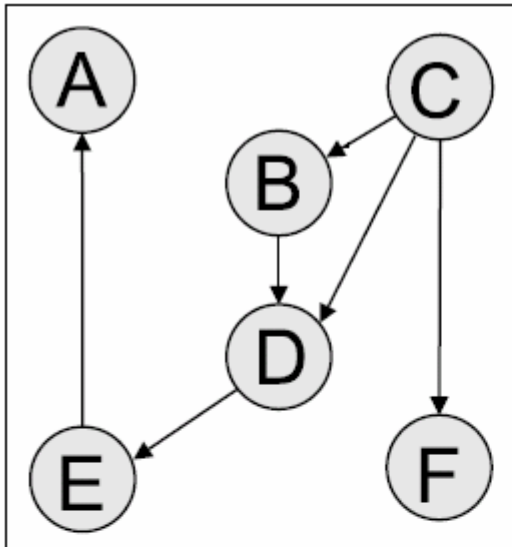
2) Orphan the nodes involved

3) Constrained resampling of the parents



This move is reversible!

1) Select an edge



2) Orphan the nodes involved

3) Constrained resampling of the parents

Mathematical Challenge:

- Show that **condition of detailed balance** is satisfied.
- Derive the **Hastings factor** ...
- ... which is a function of various **partition functions**

Acceptance probability

$$A^\triangleright(\tilde{\mathcal{M}}|\mathcal{M}) = \min \left\{ 1, \frac{N^\dagger}{\tilde{N}^\dagger} \cdot \frac{Z^*(X_i|\mathcal{M}_\ominus, X_j)}{Z^*(X_j|\tilde{\mathcal{M}}_\ominus, X_i)} \cdot \frac{Z(X_j|\mathcal{M}_\oplus)}{Z(X_i|\tilde{\mathcal{M}}_\oplus)} \right\}$$

$$Z(X_n|\mathcal{M}) := \sum_{\pi: \delta(\mathcal{M}^{X_n \leftarrow \pi})=1} \exp(\psi[X_n, \pi|\mathcal{D}])$$

$$Z^*(X_n|\mathcal{M}, X_m) := \sum_{\substack{\pi: \delta(\mathcal{M}^{X_n \leftarrow \pi})=1 \\ X_m \in \pi}} \exp(\psi[X_n, \pi|\mathcal{D}])$$

N^\dagger is the number of edges in \mathcal{M}

\tilde{N}^\dagger is the number of edges in $\tilde{\mathcal{M}}$

1
2
3
4 **Improving the structure MCMC sampler for Bayesian**
5 **networks by introducing a new edge reversal move**
6

7
8 **Marco Grzegorzcyk · Dirk Husmeier**
9

10
11
12
13
14 Received: 27 March 2007 / Revised: 11 January 2008 / Accepted: 28 March 2008
15 Springer Science+Business Media, LLC 2008
16
17

DRAFT

Evaluation

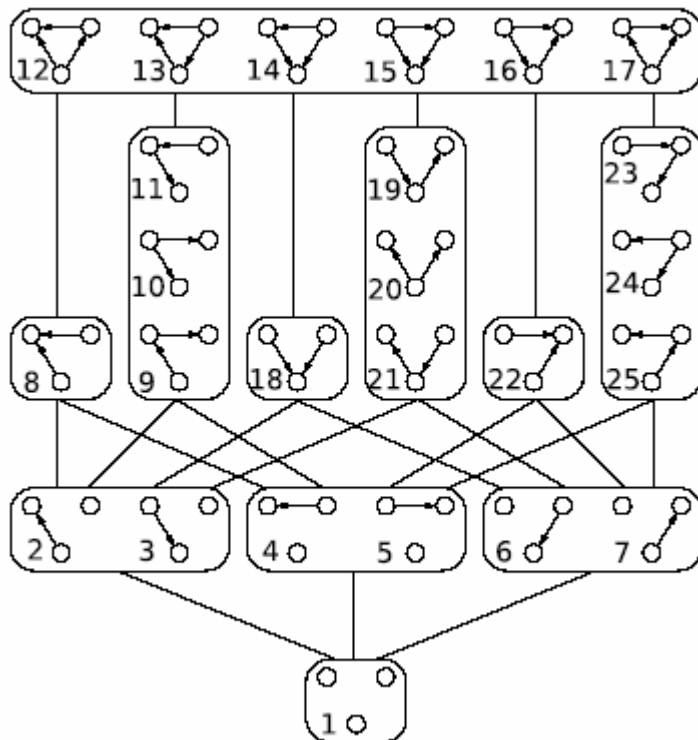
- Does the new method avoid the **bias** intrinsic to order MCMC?
- How do **convergence and mixing** compare to structure and order MCMC?
- What is the effect on the **network reconstruction** accuracy?

Results

- **Analytical comparison of the convergence properties**
- Empirical comparison of the convergence properties
- Evaluation of the systematic bias
- Molecular regulatory network reconstruction with prior knowledge

Analytical comparison of the convergence properties

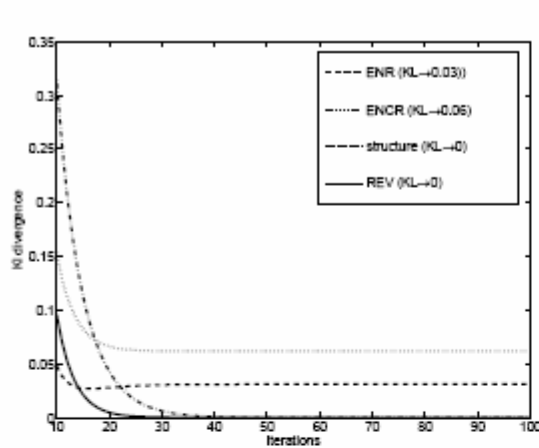
- Generate data from a noisy XOR
- Enumerate all 3-node networks



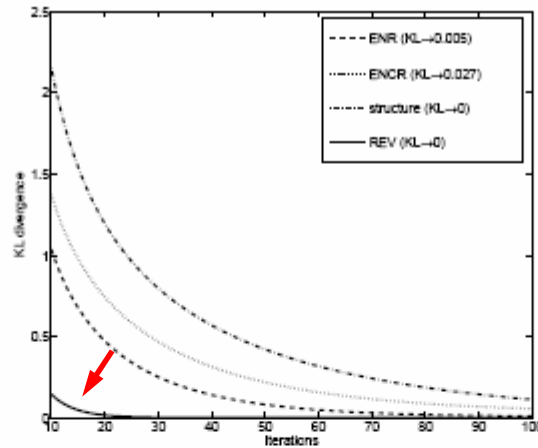
Analytical comparison of the convergence properties

- Generate data from a noisy XOR
- Enumerate all 3-node networks
- Compute the posterior distribution \mathbf{p}°
- Compute the Markov transition matrix \mathbf{A} for the different MCMC methods
- Compute the Markov chain $\mathbf{p}(t+1) = \mathbf{A} \mathbf{p}(t)$
- Compute the (symmetrized) KL divergence $KL(t) = \langle \mathbf{p}(t), \mathbf{p}^\circ \rangle$

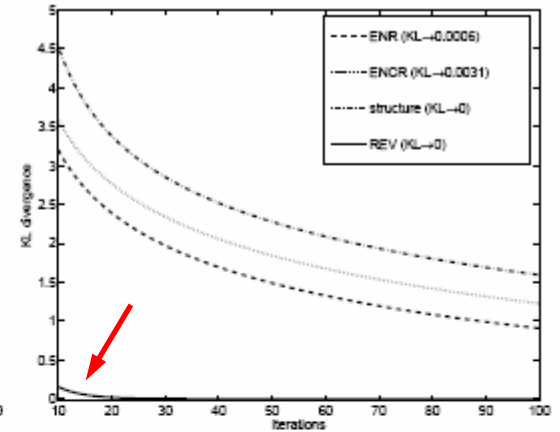
Solid line: REV-MCMC. Other lines: structure MCMC and different versions of inclusion-driven MCMC



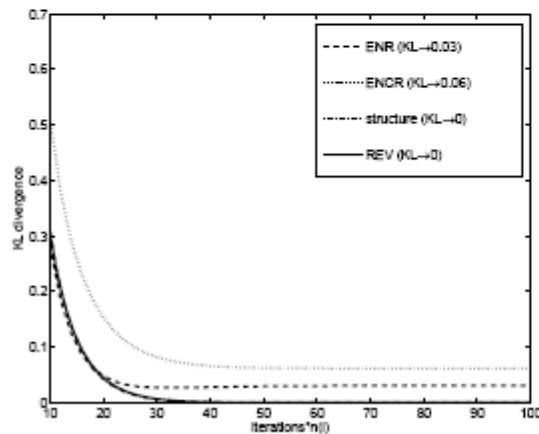
(a) $N=4$



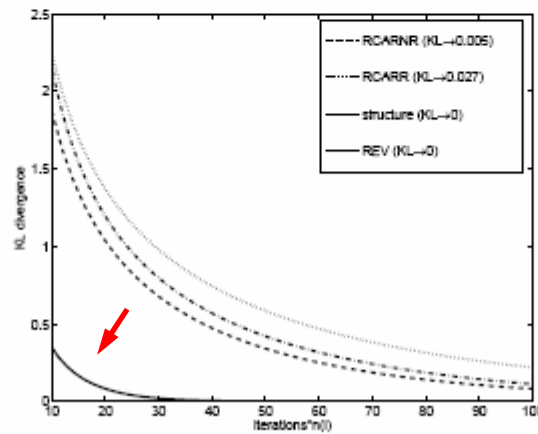
(b) $N=100$



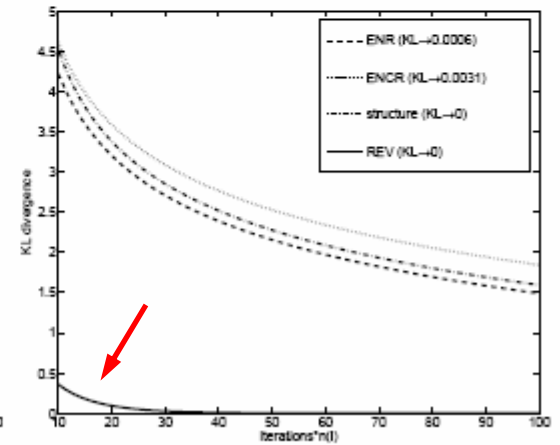
(c) $N=10000$



(d) $N=4$, adjusted



(e) $N=100$, adjusted



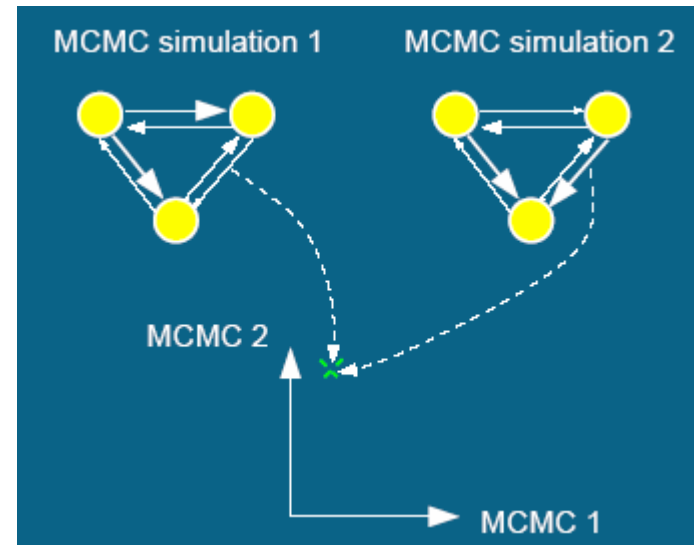
(f) $N=10000$, adjusted

Results

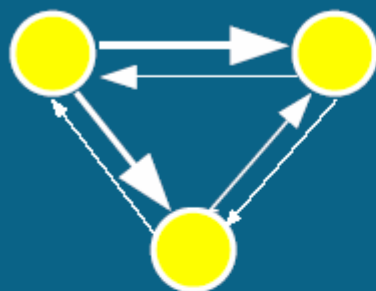
- Analytical comparison of the convergence properties
- **Empirical comparison of the convergence properties**
- Evaluation of the systematic bias
- Molecular regulatory network reconstruction with prior knowledge

Empirical comparison of the convergence and mixing properties

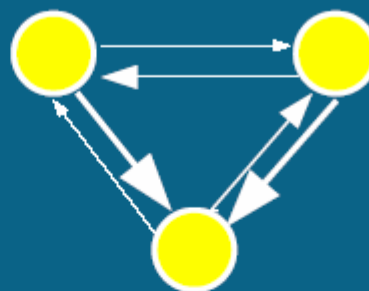
- Standard benchmark data:
Alarm network (Beinlich et al. 1989) for monitoring patients in intensive care
- 37 nodes, 46 directed edges
- Generate **data sets** of **different size**
- **Compare** the three MCMC algorithms under the same computational costs
 - **structure MCMC** (1.0E6)
 - **order MCMC** (1.0E5)
 - **REV-MCMC** (1.0E5)



MCMC simulation 1



MCMC simulation 2



MCMC 2



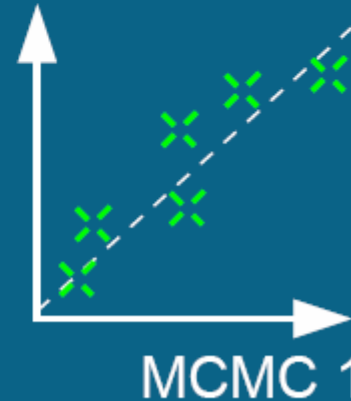
T infinite

MCMC 2



T too short

MCMC 2

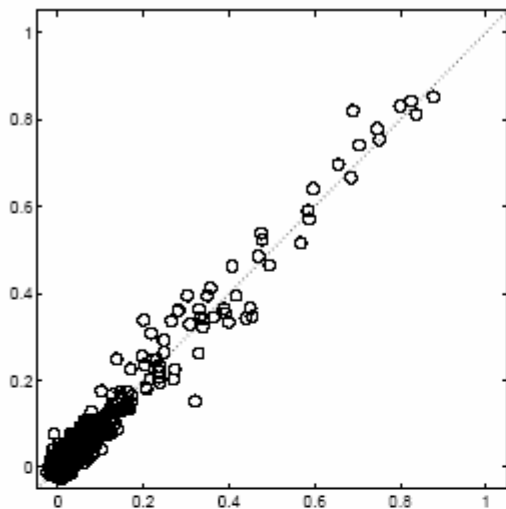


T long enough

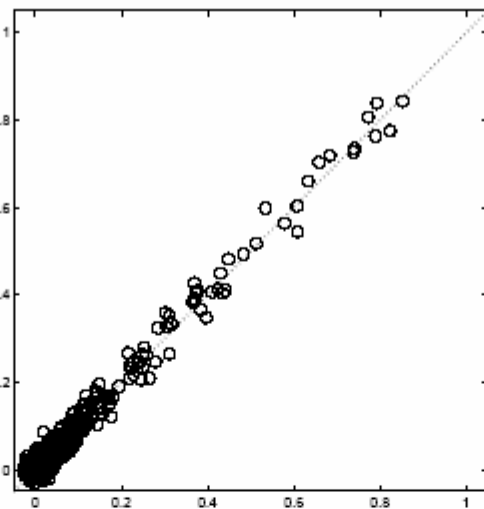
Structure MCMC

NEW

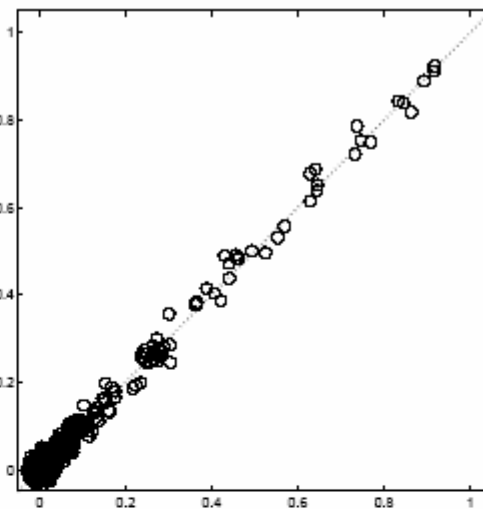
Order MCMC



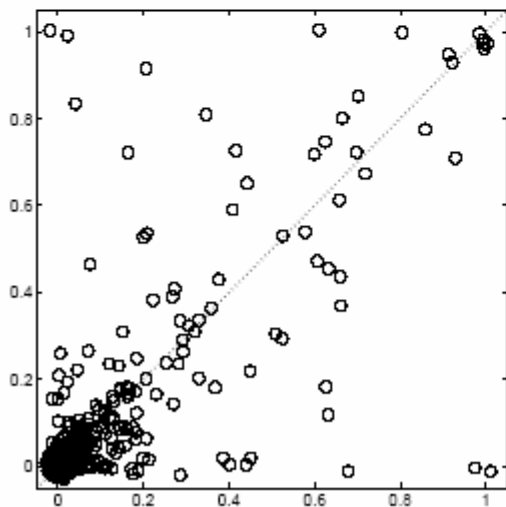
(a) Structure $m=25$



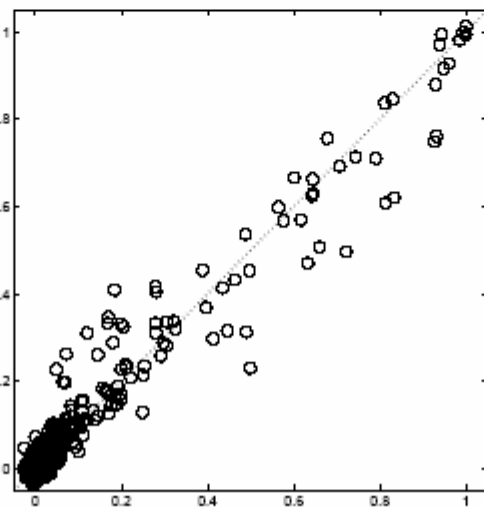
(b) REV $m=25$



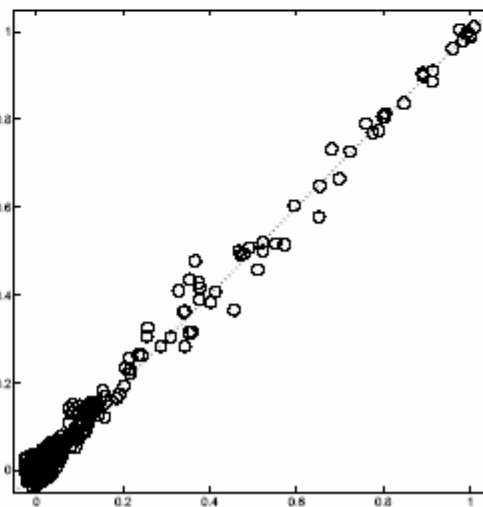
(c) Order $m=25$



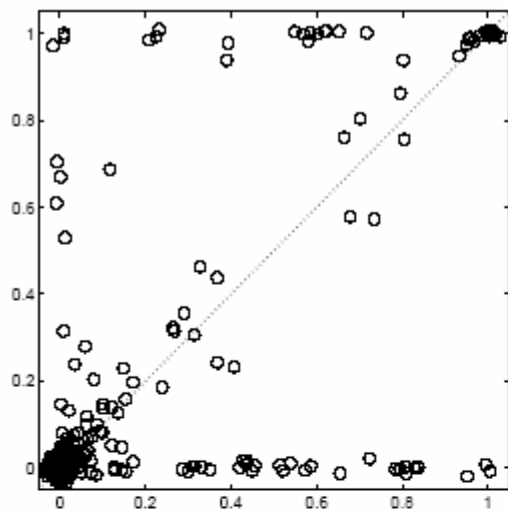
(d) Structure $m=50$



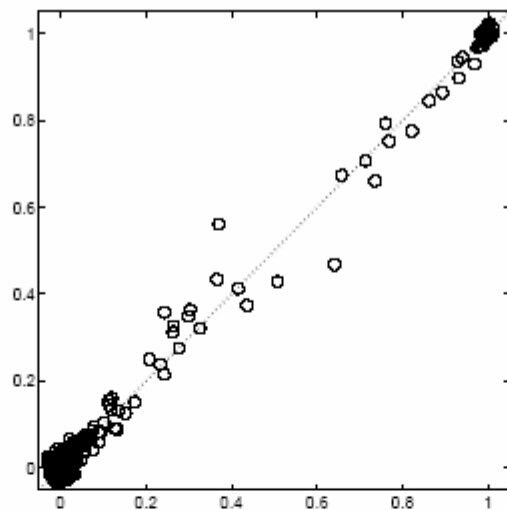
(e) REV $m=50$



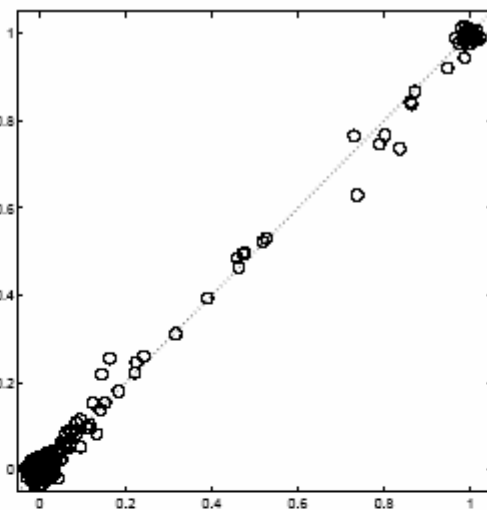
(f) Order $m=50$



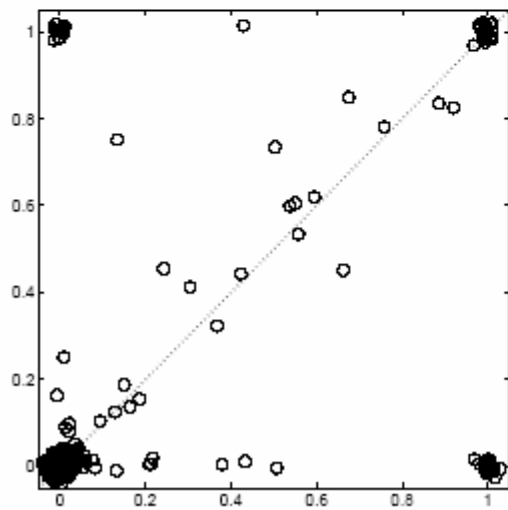
(a) Structure $m=250$



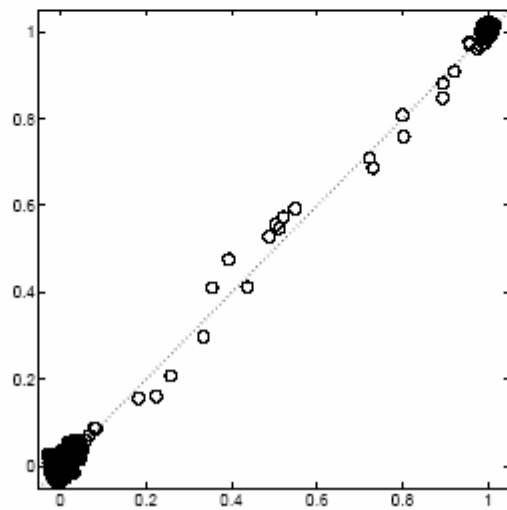
(b) REV $m=250$



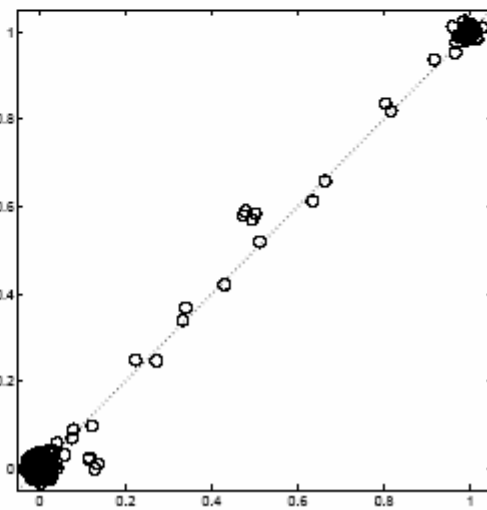
(c) Order $m=250$



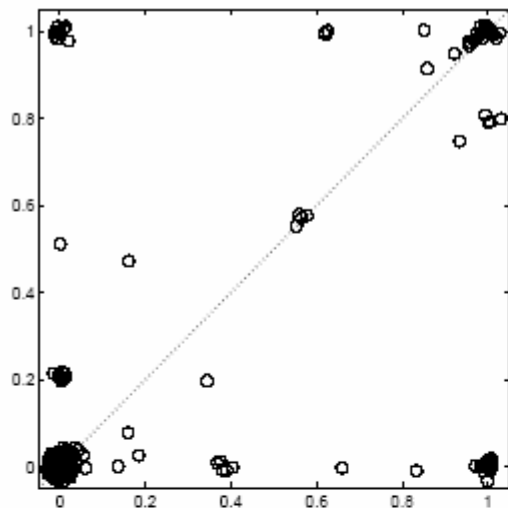
(d) Structure $m=500$



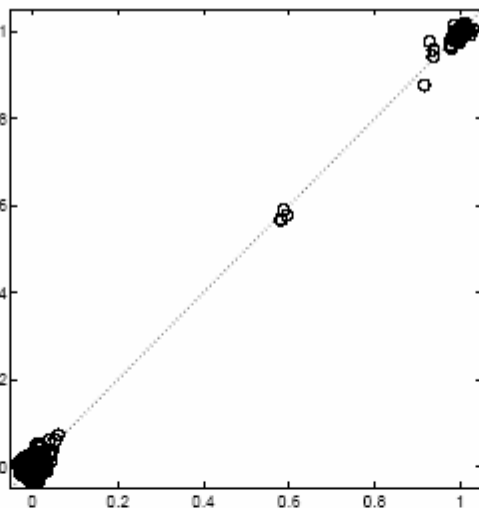
(e) REV $m=500$



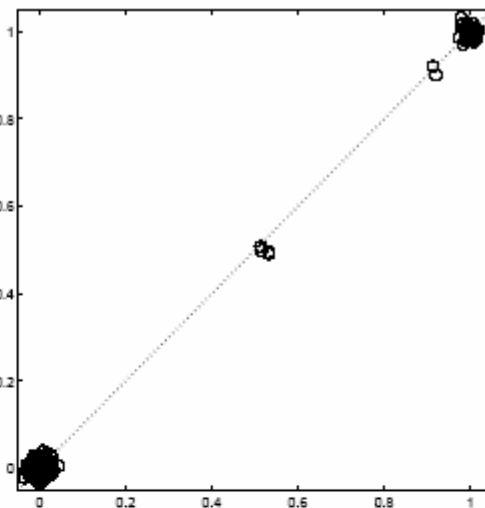
(f) Order $m=500$



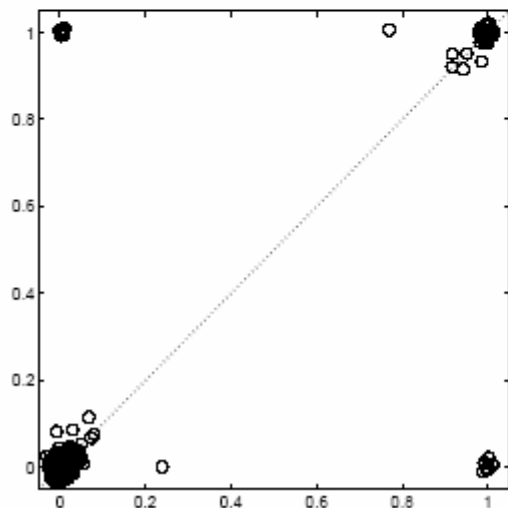
(g) Structure $m=750$



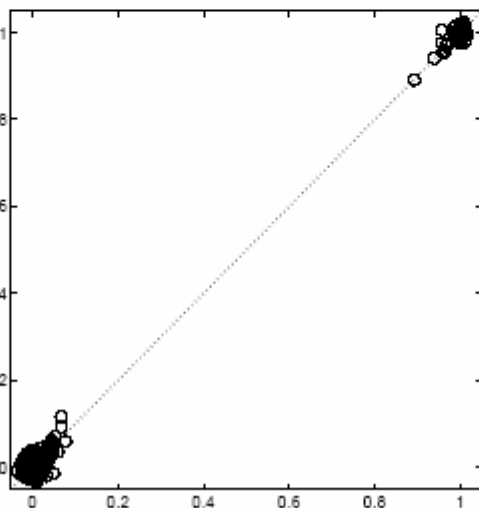
(h) REV $m=750$



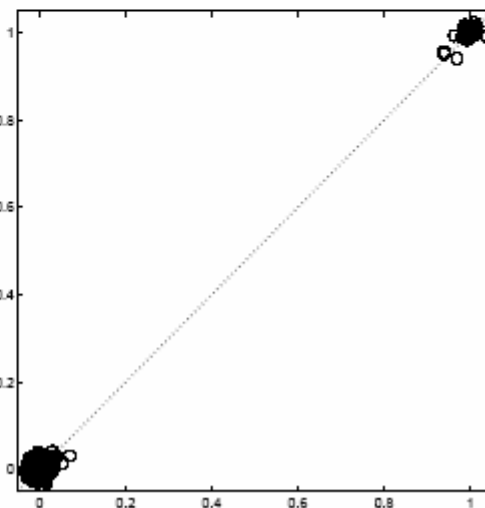
(i) Order $m=750$



(j) Structure $m=1000$



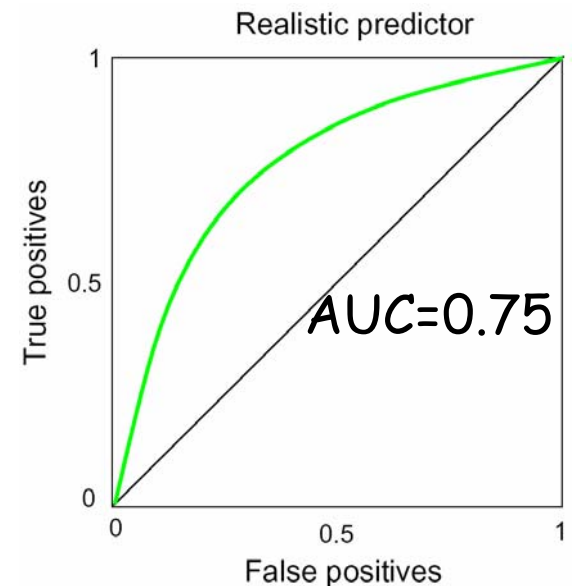
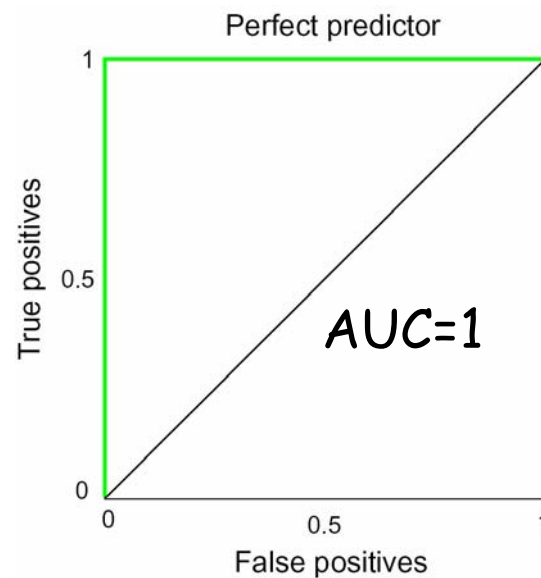
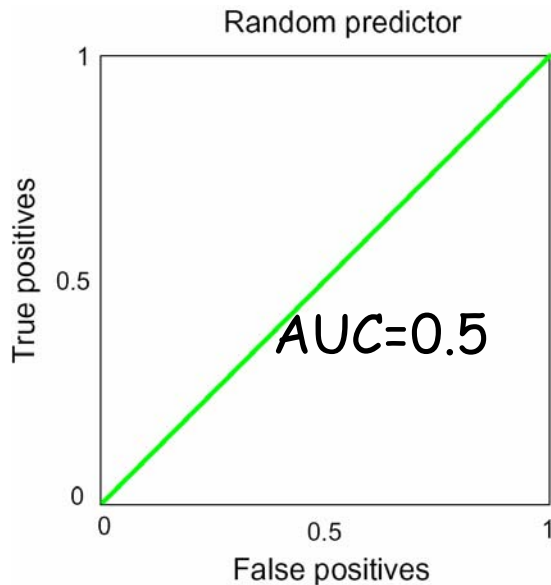
(k) REV $m=1000$

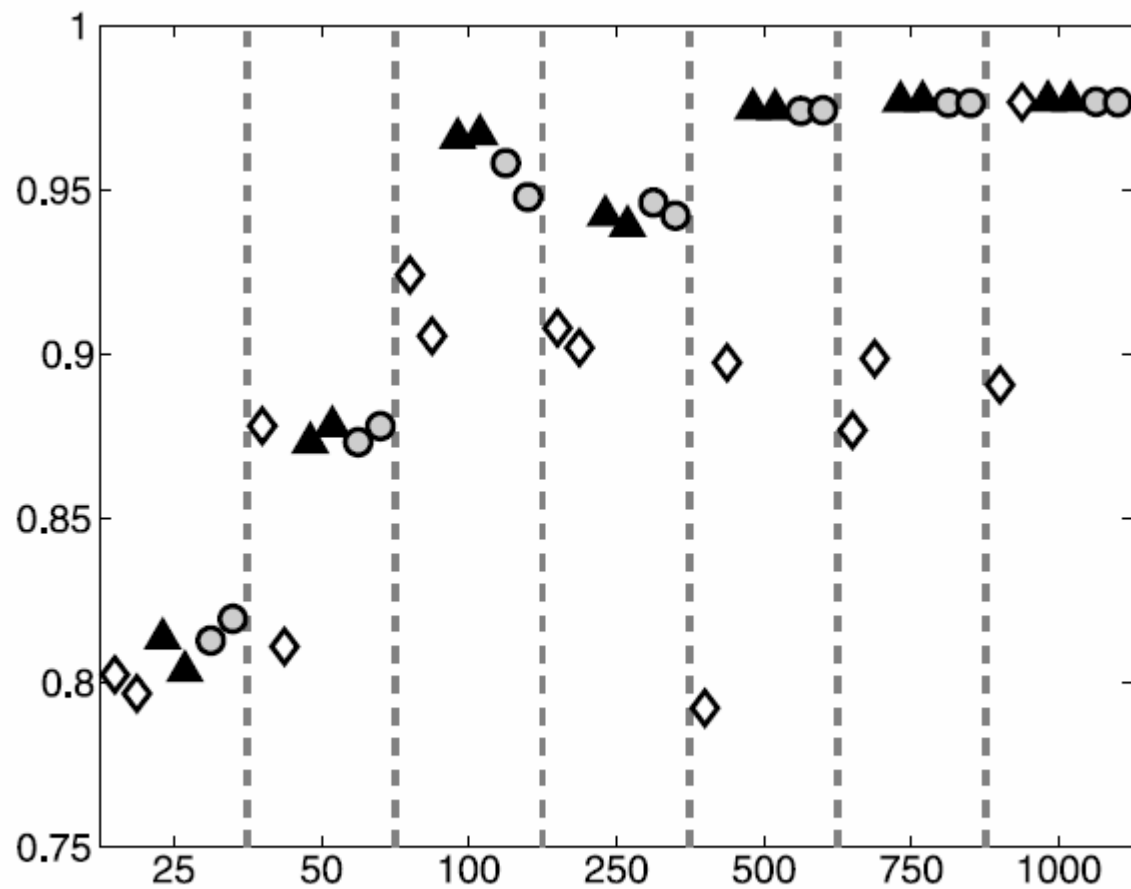


(l) Order $m=1000$

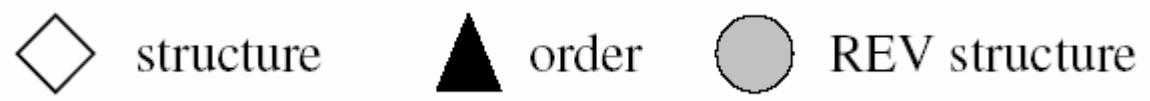
What are the implications for network reconstruction ?

ROC curves
Area under the ROC curve
(AUROC)





(a) AUROC₁



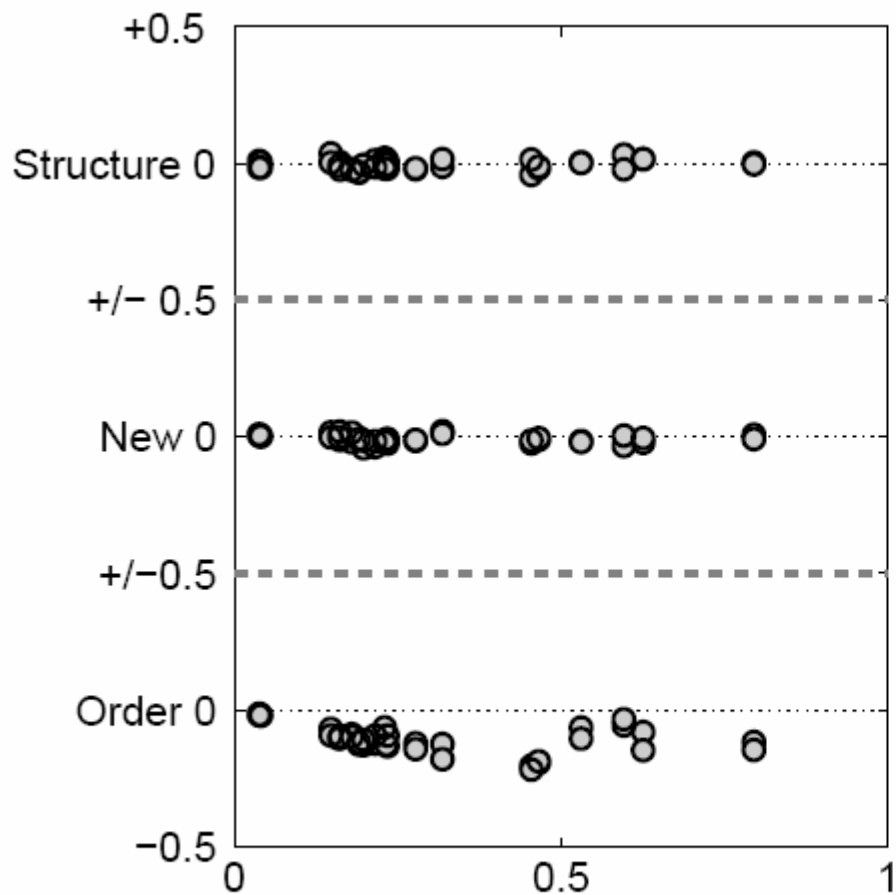
Results

- Analytical comparison of the convergence properties
- Empirical comparison of the convergence properties
- **Evaluation of the systematic bias**
- Molecular regulatory network reconstruction with prior knowledge

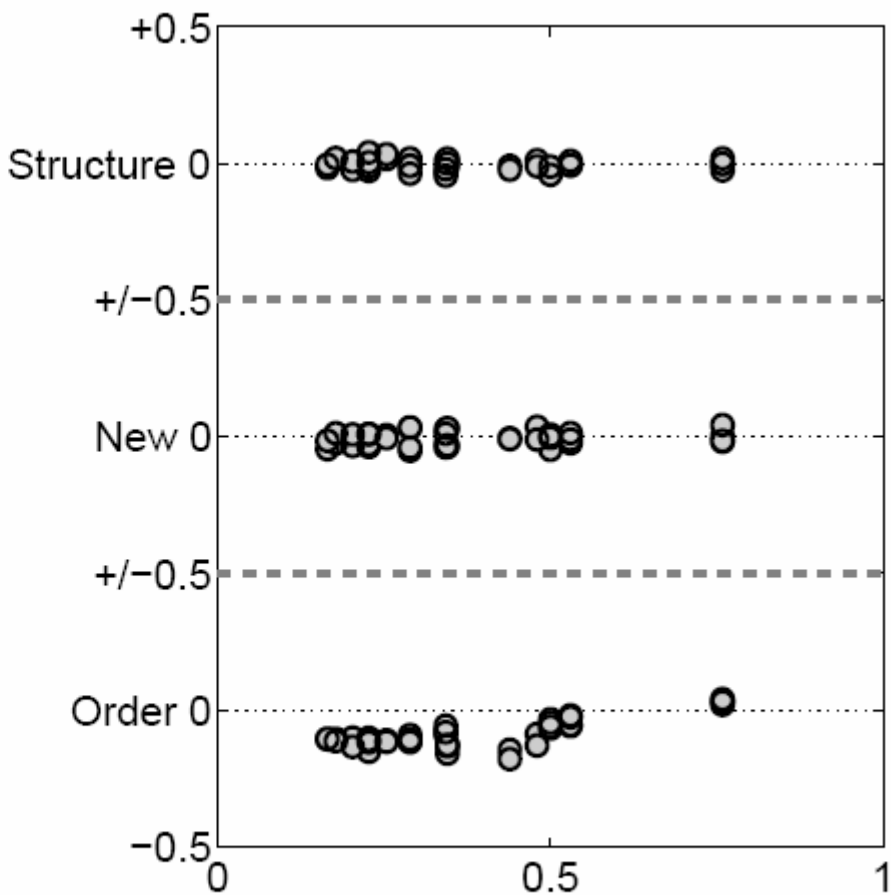
Evaluation of the systematic bias using standard benchmark data

- Standard machine learning benchmark data: FLARE and VOTE
- **Restriction to 5 nodes** → complete enumeration possible ($\sim 1.0E4$ structures)
- The **true posterior probabilities** of edge features can be computed
- **Compute the difference** between the true scores and those obtained with MCMC

Deviations between true and estimated directed edge feature posterior probabilities

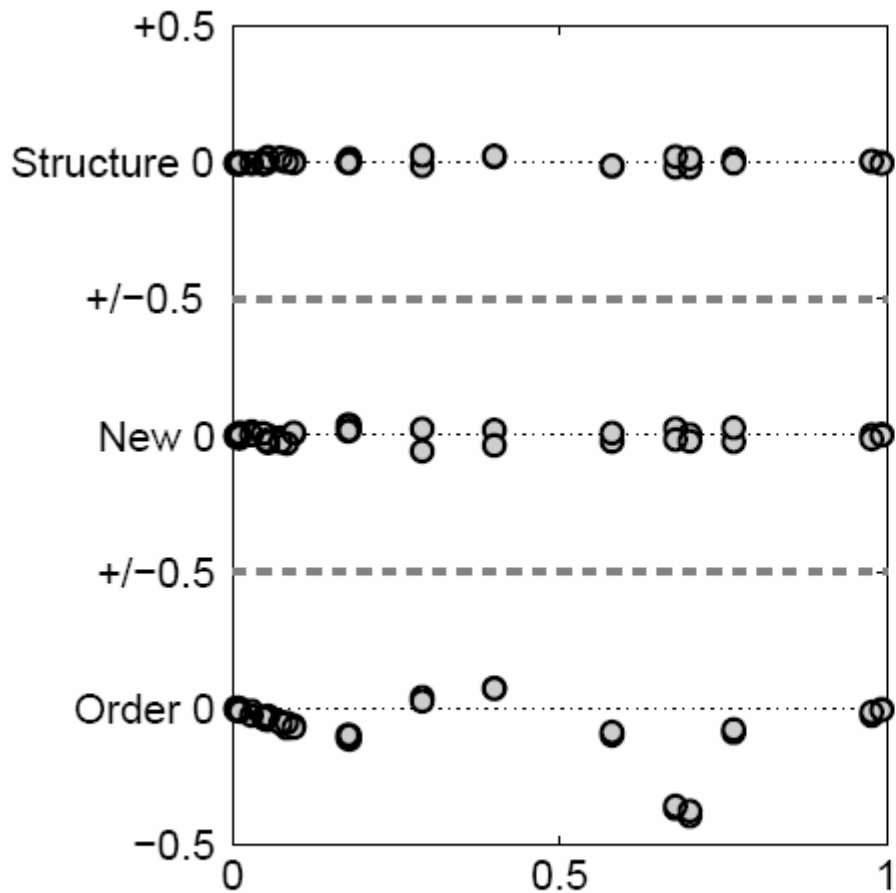


(a) FLARE $m=10$

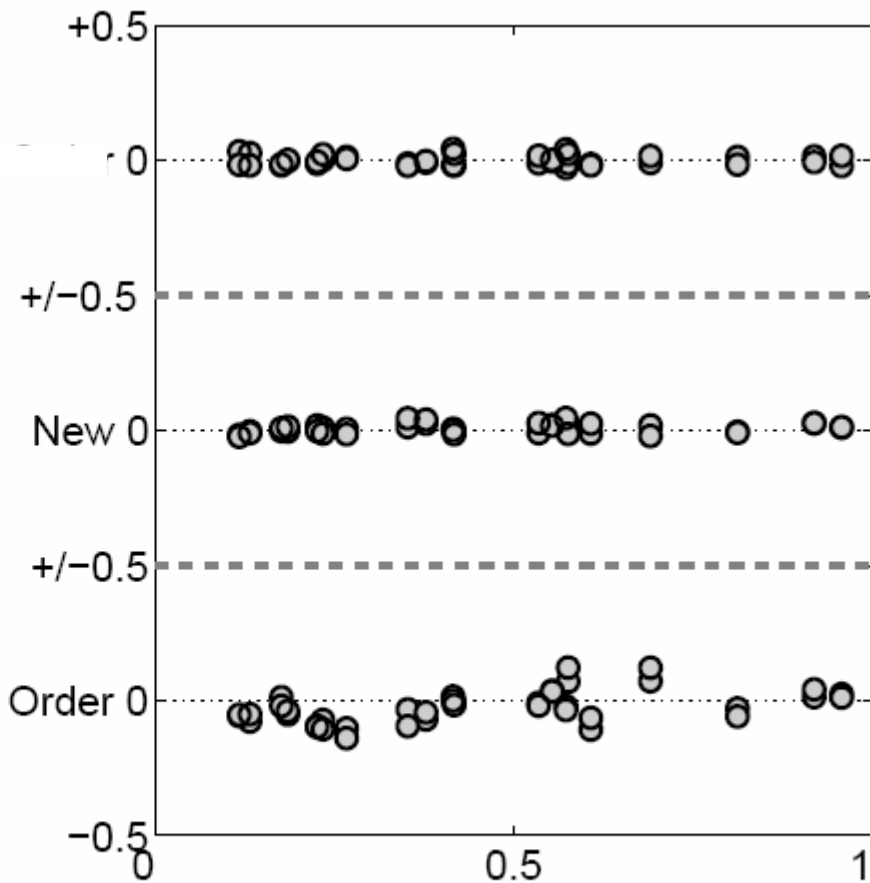


(b) VOTE $m=10$

Deviations between true and estimated directed edge feature posterior probabilities



(e) FLARE $m=50$

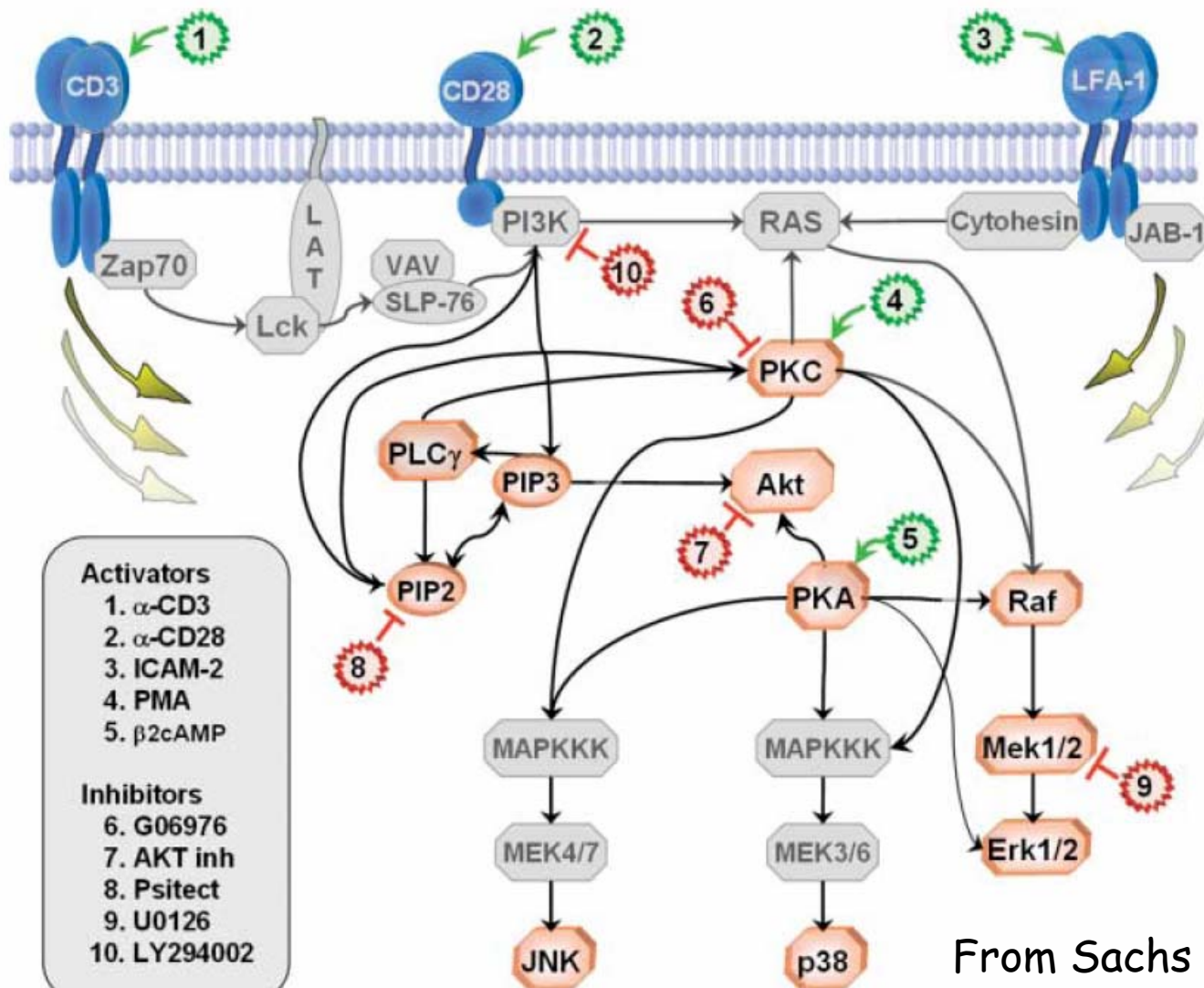


(f) VOTE $m=50$

Results

- Analytical comparison of the convergence properties
- Empirical comparison of the convergence properties
- Evaluation of the systematic bias
- **Molecular regulatory network reconstruction with prior knowledge**

Raf regulatory network



From Sachs et al Science 2005

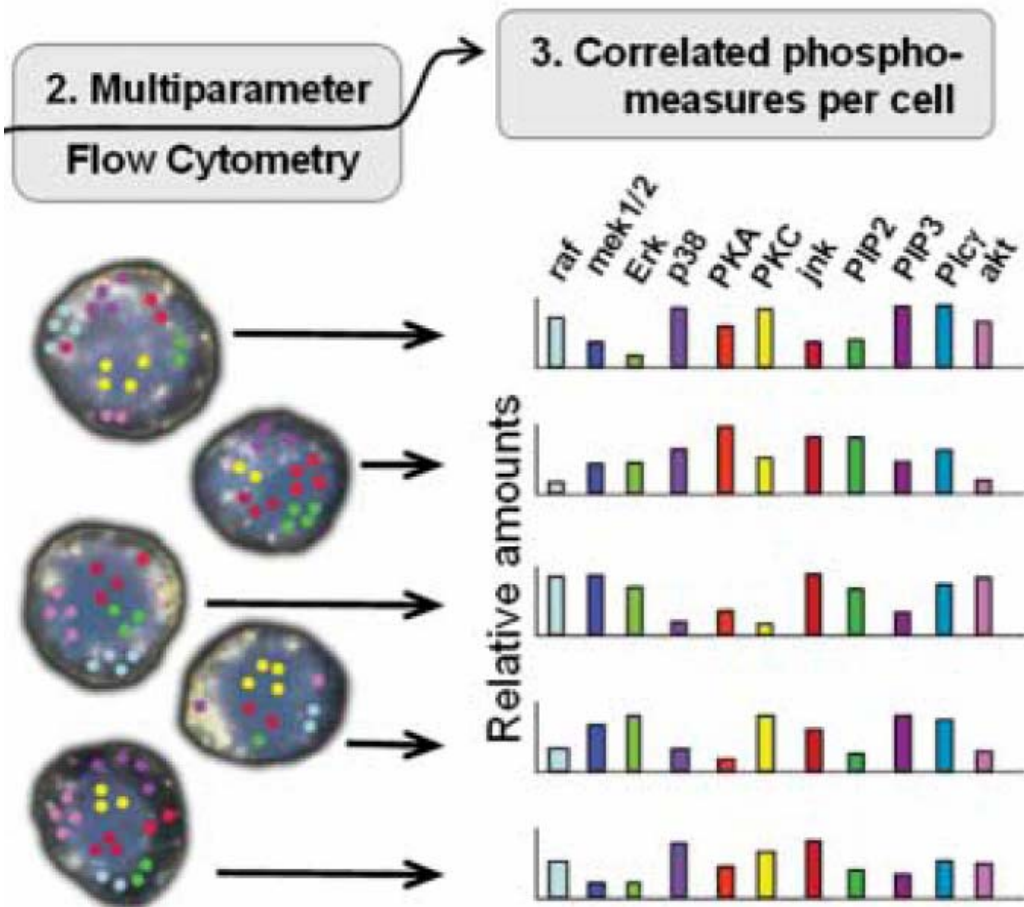
Raf signalling pathway

- Cellular signalling network of **11 phosphorylated proteins** and phospholipids in human immune systems cell
- Deregulation → carcinogenesis
- Extensively studied in the literature → **gold standard network**

Data

Prior knowledge

Flow cytometry data



Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data

Karen Sachs,^{1*} Omar Perez,^{2*} Dana Pe'er,^{3*}
Douglas A. Lauffenburger,^{1†} Garry P. Nolan^{2†}

- Intracellular multicolour flow cytometry experiments: **concentrations of 11 proteins**
- **5400 cells** have been measured under 9 different cellular conditions (cues)
- **Downsampling** to 10 & 100 instances (5 separate subsets): **indicative of microarray experiments**

Data

Prior knowledge

Deviation between the network G
and the prior knowledge B :

$$E(G) = \sum_{i,j=1}^N |B_{i,j} - G_{i,j}|$$

“Energy”

Graph: $\epsilon \in \{0,1\}$

Prior knowledge: $\epsilon \in [0,1]$

Prior distribution over networks

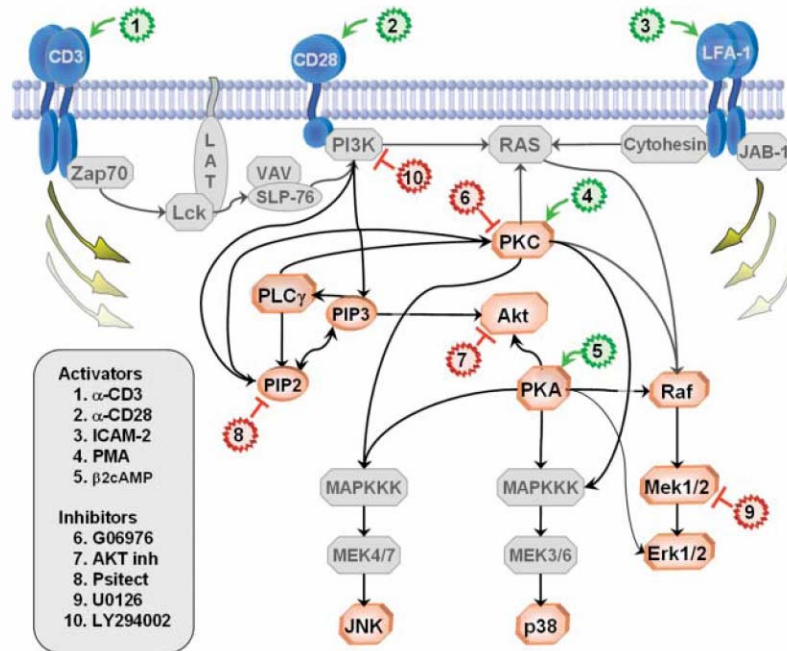
$$P(G|\beta) = \frac{e^{-\beta E(G)}}{Z(\beta)}$$

Hyperparameter

$$Z(\beta) = \sum_{G \in \mathcal{G}} e^{-\beta E(G)}$$

Prior knowledge

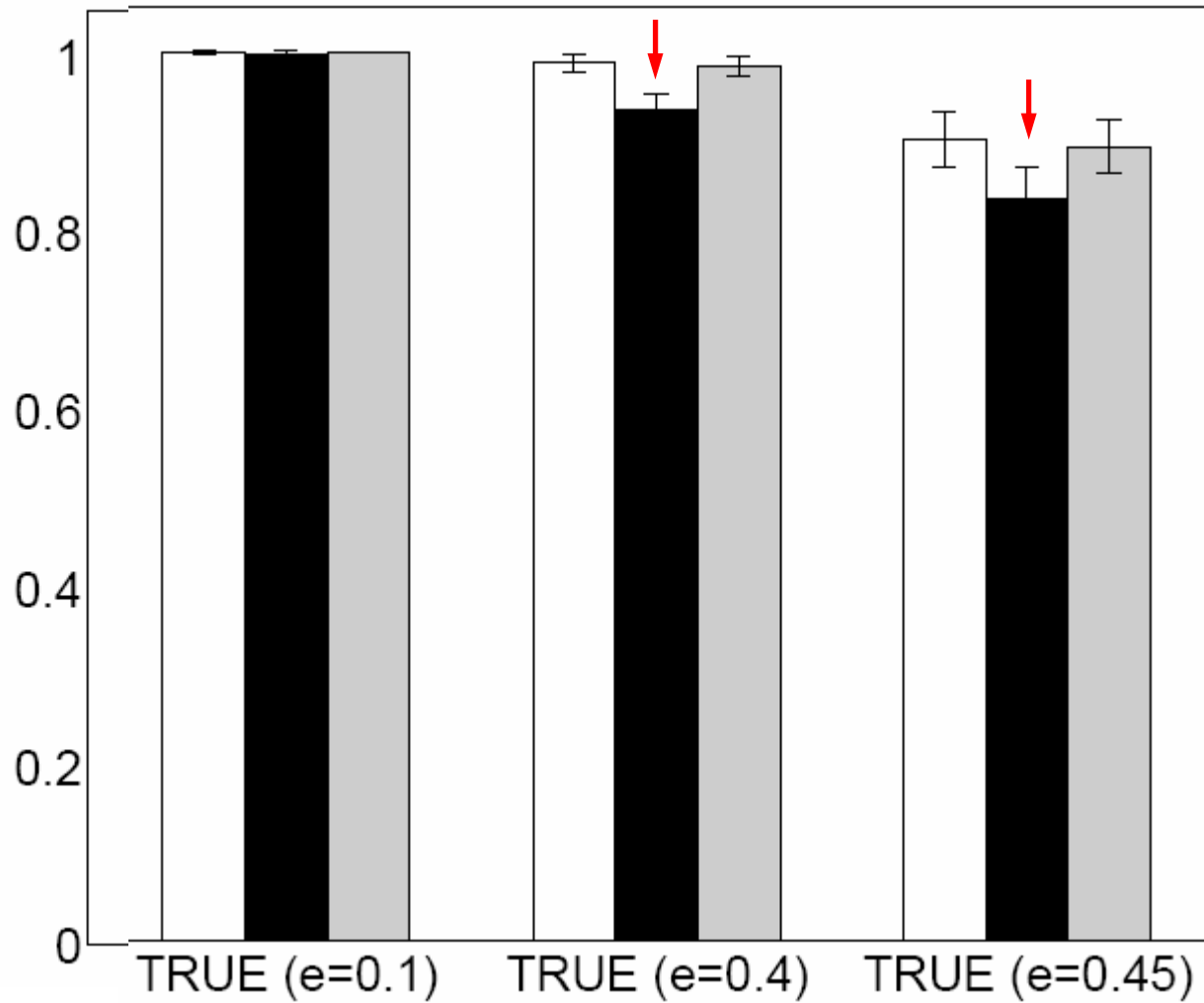
Sachs et al.



B_{ij}

Edge	Non-edge
0.9	0.1
0.6	0.4
0.55	0.45

AUROC scores



White: structure MCMC. Black: order MCMC. Grey: REV-structure

Conclusions

- The new method **avoids** the **bias** intrinsic to order MCMC.
- Its **convergence and mixing** are similar to order MCMC; both methods outperform structure MCMC.
- We can get an improvement over order MCMC when using **explicit prior knowledge**.

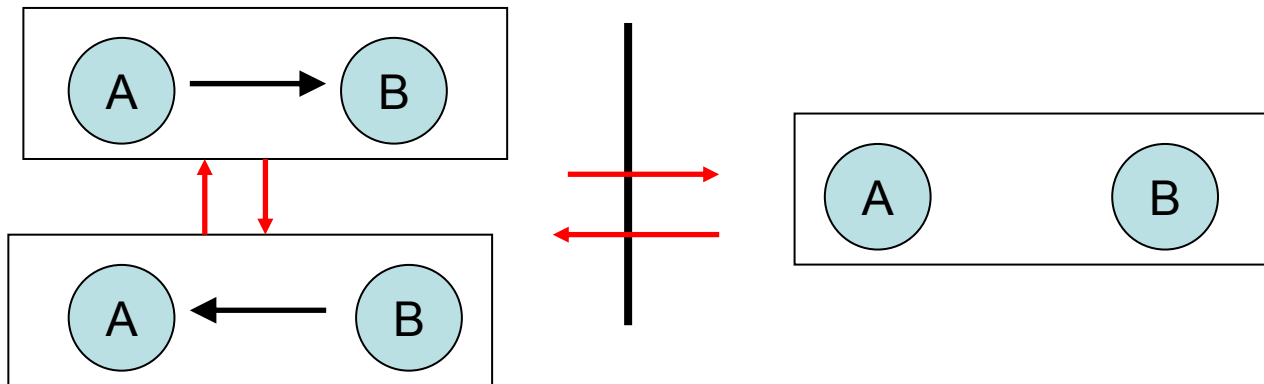
Thank you!



Any questions?

Ergodicity

- The new move is **reversible** but ...
- ... not **irreducible**



- Theorem: A **mixture** with an **ergodic transition kernel** gives an ergodic Markov chain.
- **REV-MCMC**: at each step **randomly switch** between a conventional **structure MCMC** step and the **proposed new move**.