



Structure Learning With a Small Amount of Data

Petr Šimeček

European Center for Medical Informatics, Statistics and Epidemiology,
Institute of Computer Science AS CR, Prague, Czech Republic

The structure learning could be viewed as a data-mining technique extracting hidden probabilistic dependencies between observed random variables. It is theoretically proved (in Chickering [1]) and practically verified that if the number of cases in a dataset is large enough, then it is feasible to derive the underlying dependency graph with a probability close to one. For example, models consisting of ten binary variables has been discussed in Janžura, Nielsen [2] and there has been needed a million of observations to get sufficiently precise results.

However, in medicine databases usually contain only a few hundred patients (maybe thousands in extreme cases). The question therefore arises how far is the results of these algorithms trust-able and what is a discrepancy between our forecasts and reality. This poster compares behavior of structure learning algorithms (for more details see Castillo E. et al. [3] or Neapolitan [4]) and a simple heuristic implemented in MUDIM package [5] in such situation.

Necessary Basics of Theory

A Bayesian network (BN) representing joint distribution P of random variables X_1, X_2, \dots, X_n consists of an acyclic directed graph G (dependency structure) with vertex set $\{1, 2, \dots, n\}$ and of a collection of conditional probability distributions (CPDs)

$$\{P_i; P_i = P(X_i | (X_j, j \in pa(i)))\}$$

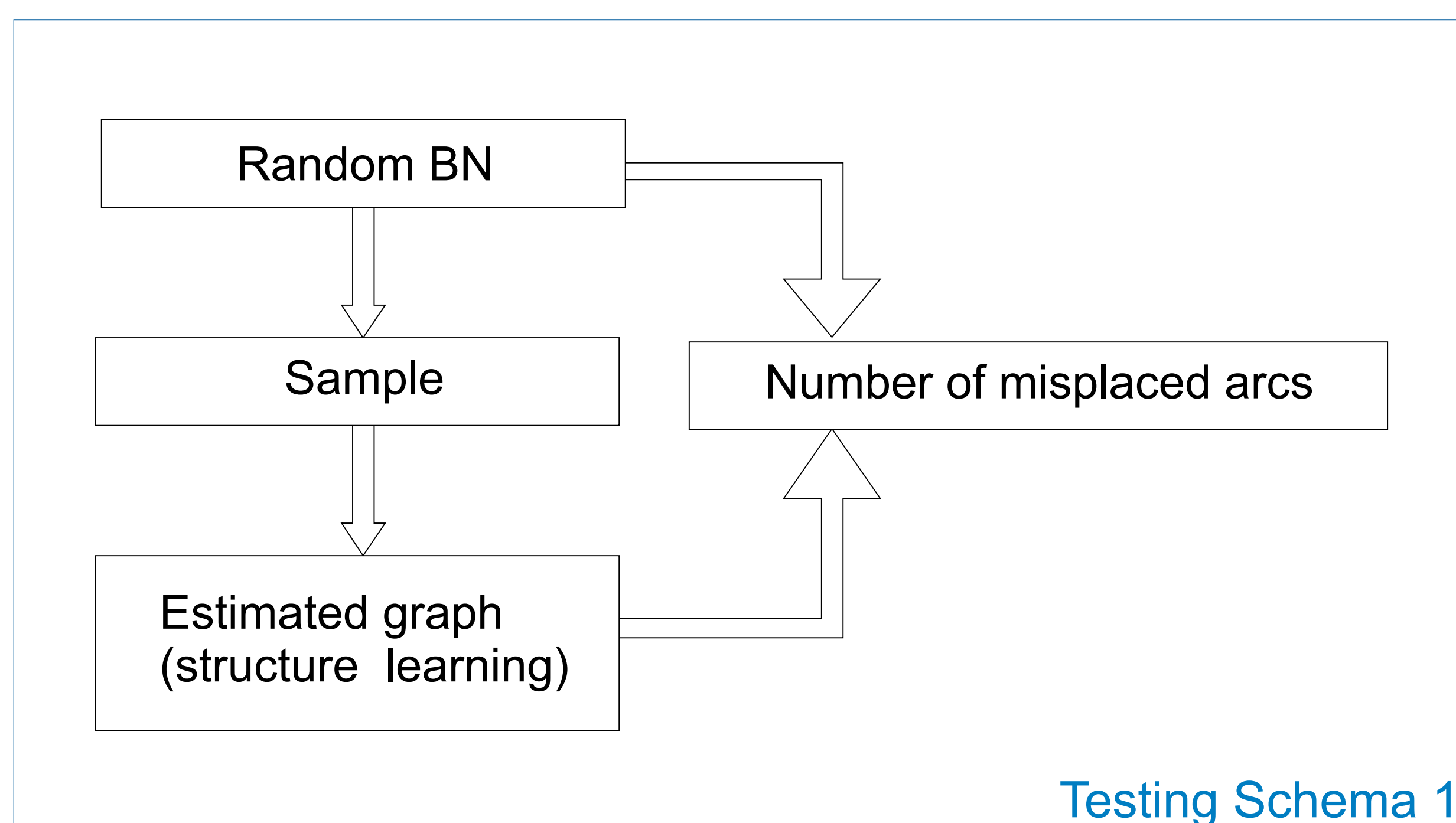
where $pa(i)$ denotes a set of parents of vertex i in G . The joint probability distribution is assumed to factorize as follows

$$P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \prod_{i=1}^n P(X_i = x_i | (X_j)_{j \in pa(i)} = (x_j)_{j \in pa(i)})$$

The goal to select the true dependency model based on observed data sample will be referred here as so called **structure learning**.

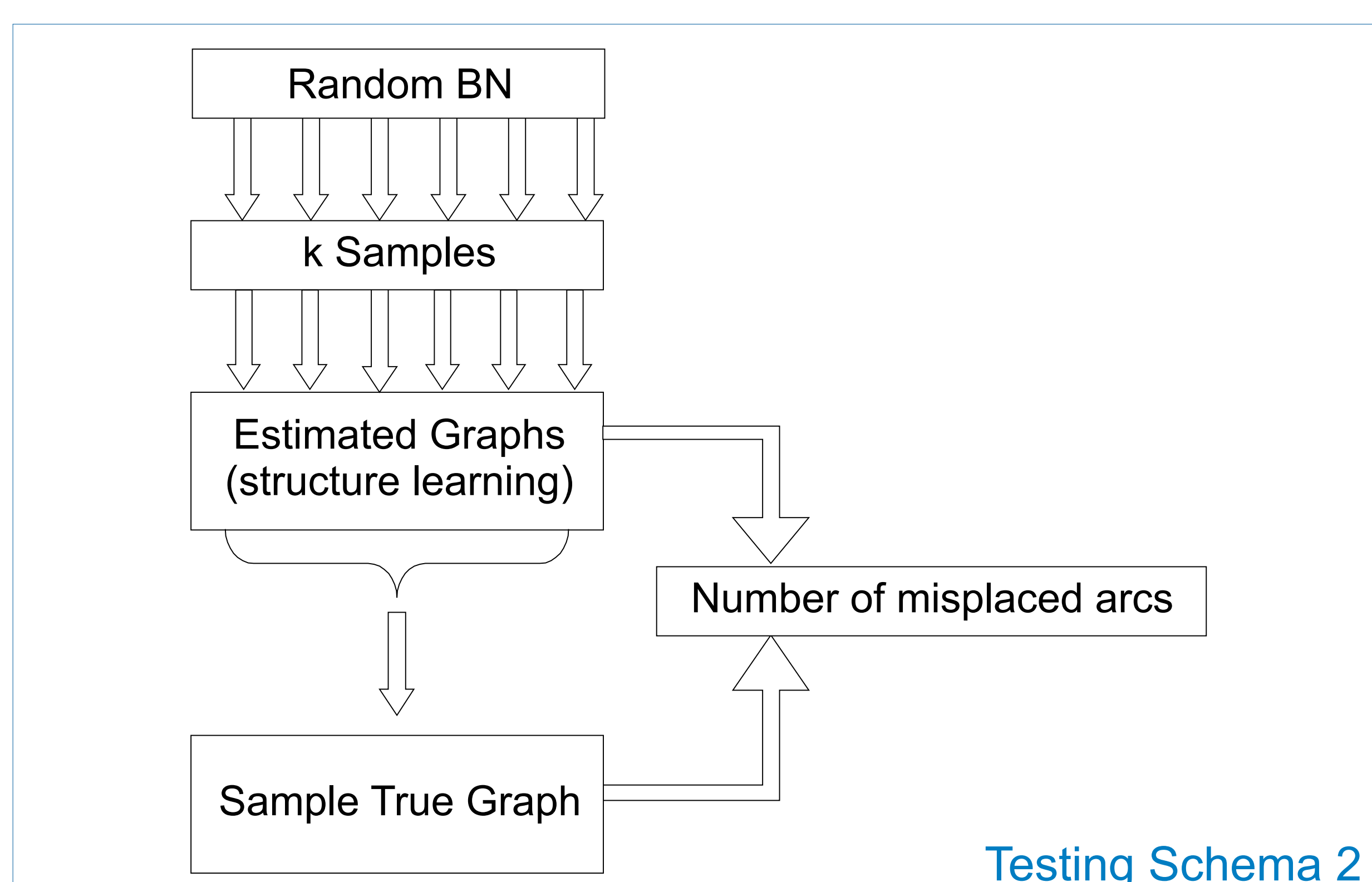
Testing Schemas

First, it is randomly uniformly chosen a dependency graph and CPDs, then a sample of this BN was generated (sample size 200). Second, it is selected dependency graph based on that sample and this graph was compare to the original (true) one. Finally, the number of misplaced arcs was recorded.

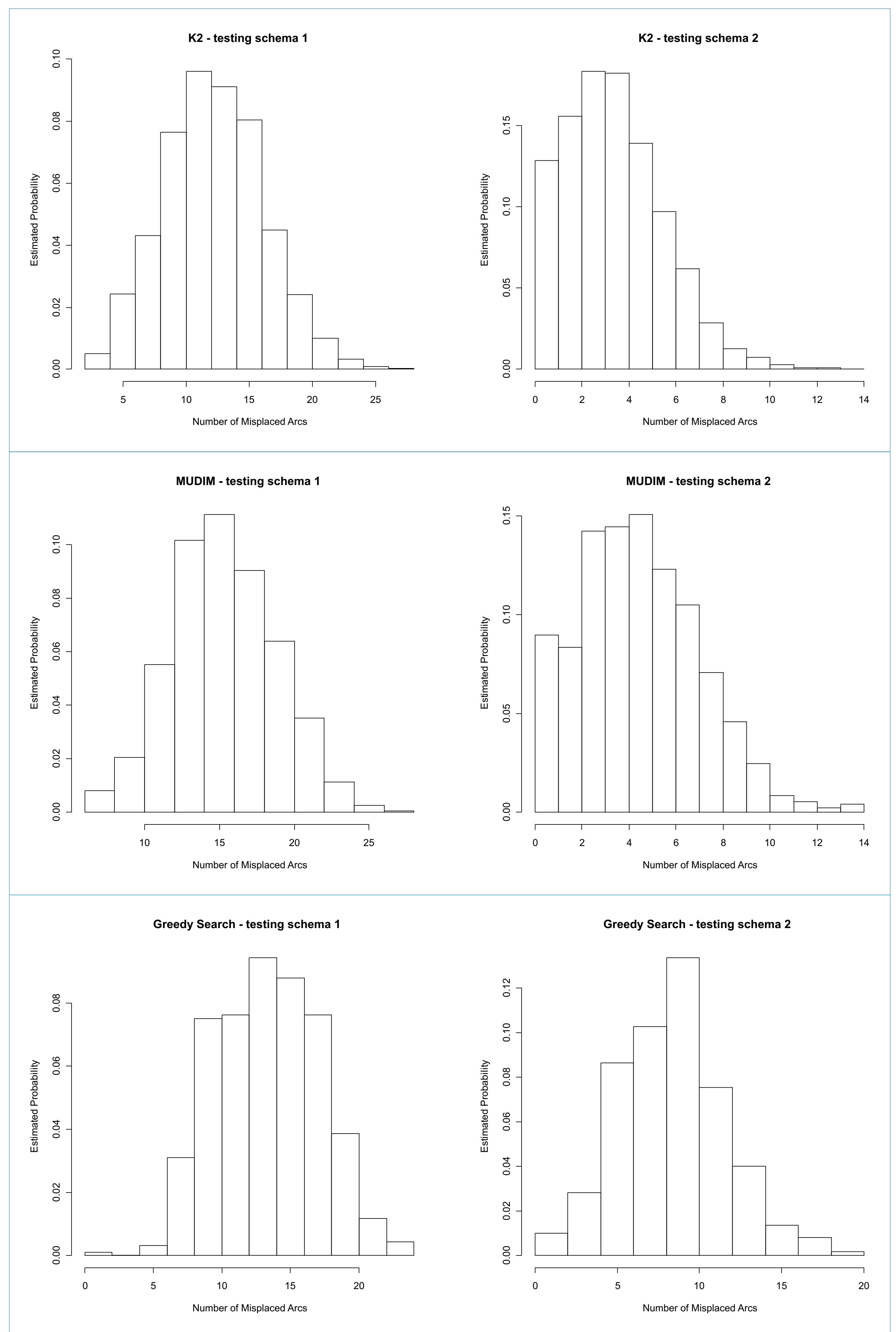


The reader can object that mistakes in selection can be partly caused by the way of generating networks. If CPDs are generated from continuous uniform distribution, there is zero probability that some dependencies from the graph do not occur. However some dependencies can be relatively weak. For that reason let us consider a slight modification of the previous algorithm:

In each step, we again generate a dependency graph and CPDs, we generate k samples (in this simulation study $k=11$) of this network and dependency graph is selected for each of these samples. These graphs are not compared with the original (true) one but with the "sample-true" graph including an edge (i,j) if and only if (i,j) is included in the majority of k selected graphs.



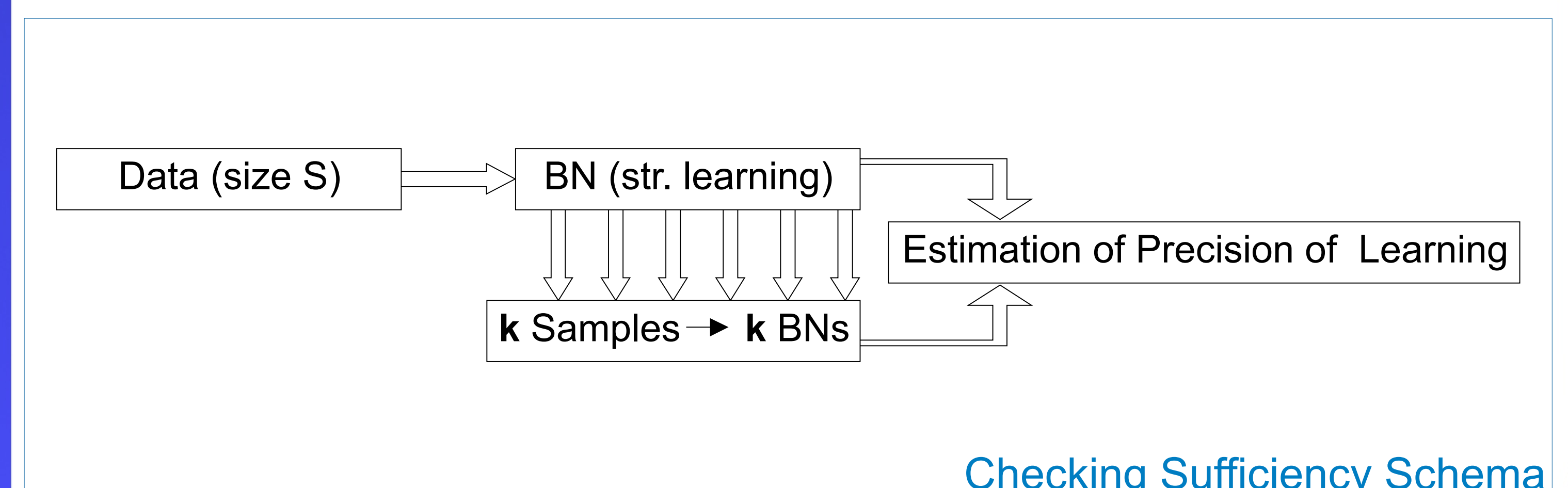
Results



The histograms of the number of mistakes (misplaced arcs) for K2 algorithm, greedy search and a heuristic implemented in MUDIM are shown above.

Conclusion

There has been shown that if the sample size is not large enough (as in medical applications) then the structural learning algorithms have a small chance to select the true model even in a case the causal ordering is known. In these cases the results of structure learning algorithms should be only viewed as approximations of the true model.



The previous method could be used to check the sufficiency of sample size as is shown on the figure above.

Acknowledgments: This work was supported by the project LN00B107 of the Ministry of Education of the Czech Republic.

References:

- [1] Chickering, M.: Optimal Structure Identification with Greedy Search, *The Journal of Machine Learning Research* 3, 507-554, 2003.
- [2] Janžura M., Nielsen J.: Method for Learning Bayesian networks from statistical data, *6th Workshop of Uncertainty Processing*, Hejnice, 2003.
- [3] Castillo et al., Expert Systems and Probabilistic Network Models, *Springer*, 1996.
- [4] Neapolitan R., Learning Bayesian Networks, *Prentice Hall*, 2003.
- [5] MUDIM, package for system R, <http://www.euromise.cz/mudim>