

# Theoretical Bounds for the Number of inferable Edges in sparse Random Networks

Frank Emmert-Streib  
Stowers Institute for Medical Research  
1000 E. 50th Street  
Kansas City, MO 64110, USA  
Email: fes@stowers-institute.org

Matthias Dehmer  
Max F. Perutz Laboratories  
Center for Integrative Bioinformatics Vienna  
Dr. Bohr Gasse 9  
A-1030 Vienna, Austria  
Email: matthias@dehmer.org

**Abstract**—The inference of a network structure from experimental data providing dynamical information about the underlying system of investigation is an important and still outstanding problem if the number of nodes within a network is not small. For example, high-throughput data from gene networks of, e.g., metabolic, signaling or transcriptional regulatory networks, provide information of thousands of genes or products thereof. Theoretically, the graph-theoretical measure d-separation provides a criteria to recover the network structure edge-by-edge by calculating the partial correlation. However, practically, for large networks it is not possible to estimate the partial correlation up to an arbitrary order because the number of possible d-separating sets grows exponential with the number of nodes in the network.

In this paper, we determine numerically theoretical bounds for the number of inferable edges in directed (possible cyclic) sparse random networks if the maximal size of d-separating sets is restricted to  $n_{max}$ . Under ideal experimental conditions these bounds correspond to the maximal precision an unknown network structure can be recovered utilizing partial correlation of order up to  $n_{max}$ .

## I. INTRODUCTION

The inference of causal network structures is an important and challenging problem. For example, in molecular biology the advent of new technological devices has led to the generation of high-throughput data allowing to monitor genes, mRNAs or proteins on a systems rather than on single molecule level. Now, the question arises if this information provided, e.g., by microarray experiments measuring the concentration of mRNAs, is sufficient to reconstruct the underlying interaction network of mRNAs and products thereof.

Mathematically, to tackle this question, one needs to define a mathematical framework defining terms like 'causal structure' to study such questions. A causal structure can be represented by a directed graph whose nodes represent the variables of the system and edges between nodes indicate a causal relationship along the direction of the edge. Important contributions, regarding the mathematical characterization of this problem, were made by Verma and Pearl [Verma and Pearl, 1988], [Pearl, 1988], [Pearl, 2000] and Spirtes, Glymour and Scheinds [Spirtes and Glymour, 1991], [Spirtes et al., 2000] who suggested algorithms to infer a causal structure from experimental data by using partial correlations if the underlying causal structure is a di-

rected, acyclic graph (DAG)  $G$ . Both algorithms, the Inductive Causation algorithm (IC) by [Verma and Pearl, 1991] and the PC algorithm by [Spirtes and Glymour, 1991], are based on the graph-theoretical measure d-separation introduced by [Verma and Pearl, 1988]. It has been proven by [Verma and Pearl, 1988] that in a DAG  $G$  node  $X$  is d-separated from node  $Y$  given node set  $S$  iff the partial correlation  $\rho_{XY.S}$  vanishes. It is remarkable to note that this relation holds for all possible model parameters as long as it is a Markovian model [Pearl, 2000]. This important theorem, relating a graph-theoretical measure to probability distributions, has been extended to directed, cyclic graphs (DCG). However, only for linear models [Spirtes et al., 1998].

In this paper we investigate the question what is the maximal percentage of edges correctly inferable from an underlying causal structure (graph) if an algorithm is used that is solely based on partial correlations up to order  $n_{max}$ . The answer to this question is of practical importance, because it allows to estimate the percentage of the inferable structure independent of any dynamics defined on the underlying directed (possibly cyclic) graph if the graph class is known.

This paper is organized in the following way. In the next section we provide the necessary framework to study your problem. In section III we present our results from numerical simulations. This paper finishes in section IV with a summary and concluding remarks.

## II. MATHEMATICAL FRAMEWORK

In the following, we aim to estimate the percentage of edges inferable from a directed graph (possibly cyclic) with an algorithm that is solely based on partial correlations of order up to  $n_{max}$ . Due to the fact, that the partial correlation is symmetric in its arguments we do not aim to estimate the direction of the edges but only their presence or absence. The application of partial correlations to infer the structure of the underlying network has practically two drawbacks. First, for large graphs the search for a set  $S$  d-separating two nodes  $X$  and  $Y$  disjunct and not included in  $S$  can be hard, because of the combinatorial explosion of possible sets  $S$  which need to be tested for independence. The total number of tentative  $S$

sets for a graph with  $N$  nodes is

$$\sum_{|S|=0}^{N-2} \binom{N-2}{|S|} \quad (1)$$

$|S|$  is the number of nodes in set  $S$ . Second, for directed, cyclic graphs the partial correlation does not necessarily vanish for variables not directly connected in the true model. Fig. 1 depicts such a situation. Node  $X_1$  and  $X_3$  are not directly

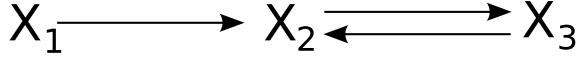


Fig. 1. A directed cyclic graph.  $X_1$  is not d-separable by any possible set from  $X_3$ .

connected, however, they are not d-separated by any of the possible sets  $\{\{\emptyset\}, \{X_2\}\}$  whereas d-separation is given by the following definition [Verma and Pearl, 1988], [Pearl, 1988]<sup>1</sup>.

**Definition 1:** Two nodes  $X$  and  $Y$  are called d-separated by set  $S$  if  $X$ ,  $Y$  and  $S$  are disjunct and every undirected path from  $X$  to  $Y$  is blocked by the nodes in set  $S$ .

**Definition 2:** A path  $w$  is called d-separated or blocked by a set of nodes  $S$  iff one of the two criteria holds

- 1) on  $w$  is a node  $s$  who is no collider and  $s \in S$
- 2) on  $w$  is a node  $s$  who is a collider and it holds  $s \notin S$  and  $de(s) \notin S$ .

Here,  $de(s)$  denotes the set of descendants of node  $s$  and a node  $X$  is called a collider if the edge of the incoming as well as the edge of the outgoing path points to  $X$ .

In this case, the inferred undirected graph based only on partial correlations, would look like Fig. 2 containing additionally an edge connecting node  $X_1$  and  $X_3$  directly, which is apparently wrong.

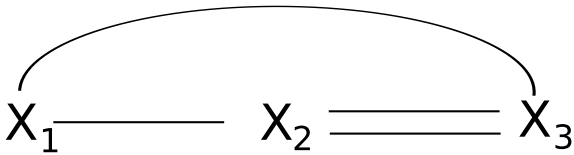


Fig. 2. Inferred structure based on partial correlations.

Because of these two limitations of algorithms using exclusively partial correlation to infer the structure of the underlying graph we aim to answer the following question. Given a directed (possibly cyclic) graph  $G$  - What is the number of edges that can be correctly inferred by using the d-separation criterion? For a given graph  $G$  with  $N$  nodes the maximal number of possible edges without self-connections is

$$N^{tot} = \binom{N}{2} = \frac{N(N-1)}{2}. \quad (2)$$

<sup>1</sup>We repeat the definitions for convenience.

Theoretically, the application of the d-separation criterion to the nodes  $X$  and  $Y$  gives the following possible results: If the nodes  $X$  and  $Y$  are d-separable then there exists a set  $S$  consisting of zero, one, two ... or  $N - 2$  disjunct nodes. Otherwise the nodes  $X$  and  $Y$  are not d-separable. If  $G$  is a DAG the number of node pairs that are not d-separable  $N_{DAG}^{nd-sep}$  corresponds to the number of node pairs  $N_{DAG}^c$  directly connected. For  $G$  a DCG this number  $N_{DCG}^{nd-sep}$  can be larger than  $N_{DCG}^c$  as demonstrated with the help of Fig. 1. Let  $p_{DAG}$  and  $p_{DCG}$  be the percentage of correctly inferred edges of a DAG or a DCG graph  $G$ . Hence, if  $G$  is a DAG  $p_{DAG} = 1$ , but if  $G$  is a DCG the number of correctly inferable edges is

$$p_{DCG} \leq \frac{N_{DCG}^{d-sep} + N_{DCG}^{nd-sep}}{N^{tot}} = 1 \quad (3)$$

with

$$p_{DCG} = \frac{N_{DCG}^{d-sep} + N_{DCG}^c}{N^{tot}} \quad (4)$$

Here  $N_{DCG}^{nd-sep} - N_{DCG}^c$  is the number of node pairs that are not directly connected, but are nevertheless not d-separable. We want to remark, that Eq. 3 follows only from theoretical considerations not including practical problems that occur in reality like, e.g., the estimation of partial correlations from small sample sizes.

The second effect that might reduce  $p_{DCG}$  results from the intractability to test all candidate sets  $S$  that could d-separate two given nodes. That means, practically, we can not test all possibilities, but have to restrict the complexity of the analysis. It has been suggested [de la Fuente et al., 2004] to calculate the partial correlation only up to order  $n$ . In practical investigations  $n$  has been chosen heuristically to one or two [de la Fuente et al., 2004], [Magwene and Kim, 2004], [Wille and Bühlmann, 2006]. Regardless, how  $n$  is chosen, it leads to the splitting of the number of d-separable node pairs in a sum of two terms given by

$$N^{d-sep} = N_{\leq n}^{d-sep} + N_{> n}^{d-sep} \quad (5)$$

whereas the first term on the rhs gives the number of node pairs that are d-separable with separating sets  $S$  with sizes  $|S| \leq n$  and the second term the number of node pairs that are d-separable with separating set sizes  $|S| > n$ . This gives us a new estimation for the percentage of inferable edges

$$p'_{DCG} = \frac{N_{\leq n}^{d-sep} + N_{DCG}^c}{N^{tot}} \quad (6)$$

So far it is unknown what quantitative effect the restriction of the order of the partial correlation on  $p'_{DCG}$  has.

In the following we ask the question what  $p'_{DCG} \leq 1$  numerically means. For example  $p'_{DCG} \in \{0, 0.1, 1\}$  represent three valid, numerical values for  $p'_{DCG}$  that are consistent with Eq. 3. However, the smaller  $p'_{DCG}$  the worse the results. In the next section we determine numerically theoretical bounds for  $p'_{DCG}$  for directed (possibly cyclic), sparse random networks.

### III. RESULTS

For the following numerical studies we use a graph class called random networks that was independently proposed by [Solomonoff and Rapoport, 1951] and [Erdős and Rényi, 1959]. A random graph is generated by connecting each possible pair of nodes with probability  $q$ . The resulting graph with  $N$  nodes has a mean degree of  $z \sim qN$  [Newman, 2003]. We use  $z$  as a control parameter, because we are only interested in sparse graphs having only a few connections per node. The reason therefore is twofold. First, your studies aim to improve the analysis of high-throughput data from molecular biology. It has been argued that the underlying networks of gene interaction or products thereof are sparsely connected [Jeong et al., 2000], [Jeong et al., 2001]. Second, the graph-theoretical measure d-separation we use requires to determine all undirected paths within a network to find the node set which blocks all paths. This is prohibitive for graphs of large order and/or high mean degree, because this leads to an exponential increase in the number of paths.

We determine the distribution of  $n$ , the size of the d-separating sets, with the following algorithm.

*Algorithm 1:* Given a directed (possibly cyclic) graph  $G$  with  $N$  nodes

- repeat until all different pairs of nodes from  $G$  are selected
  - determine all undirected paths from node  $X$  to node  $Y$
  - if  $X$  and  $Y$  are directly connected save  $n =$  'directly connected' and start with a new pair of nodes
  - if no path is found connecting  $X$  and  $Y$  save  $n =$  'not connected' and start with a new pair of nodes
    - \* test for d-separability starting with  $n = 0$  nodes and increase the number of used nodes by one up to  $n_{max}$
    - \* if a set of nodes  $S \setminus \{X, Y\}$  is found that d-separates  $X$  and  $Y$  save  $n = |S|$  otherwise save  $n =$  'not found' and start with a new pair of nodes

Algorithm 1 determines the number of d-separating sets of size 0 to  $n_{max}$ . For our simulations we set  $n_{max} = 3$ . Moreover, algorithm 1 determines the number of node pairs that are: not connected by a path (nc), directly connected (c), not d-separable up to a set size  $n_{max}$  ( $> n$ ). The symbols in brackets correspond to the symbols in Fig. 4 to 9. The normalized distributions for these values are shown in Fig. 4 to 9 for different parameter values of  $N$  and  $q$ .

Each distribution (normalized histogram) was obtained by averaging over an ensemble of 1000 networks independently generated. On the left hand side of each distribution are the results for d-separating sets of size 0 to  $n_{max} = 3$ . On the right hand side are the values for nc = 'not connected', c = 'connected' and  $> n =$  'not d-separable with set sizes up to  $n_{max}$ '.

These simulations demonstrate clearly, that the percentage of edges not correctly detectable with d-separating set sizes up

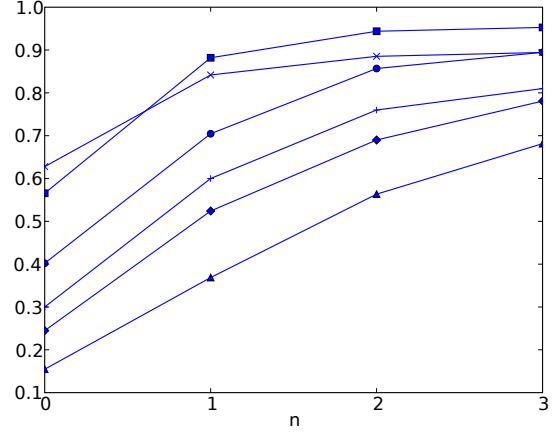


Fig. 3. Theoretical bounds for the precision of inferable network structures for sparse random networks in dependence on  $n$  (size of d-separating sets). Plus (+):  $N = 20, q = 0.15$ , cross (x):  $N = 20, q = 0.05$ , diamond (◊):  $N = 20, q = 0.0789$ , circle (o):  $N = 30, q = 0.10$ , triangle (Δ):  $N = 20, q = 0.10$  and square (□):  $N = 50, q = 0.03$ .

to size  $n_{max} = 3$  is about 5%. However, this does not mean that this is the total error. Given a network structure of a graph  $G$  we are able to distinguish between pairs of nodes that are directly connected and pairs of nodes that are not d-separable with set sizes up to  $n_{max}$ . However, it is not possible to distinguish these conditions utilizing estimators for the partial correlation coefficients from, e.g., time series data, even under ideal experimental conditions and infinite long time series, because neither the partial correlation between directly connect nodes vanishes nor the partial correlation between nodes which are *not* d-separable with set sizes up to  $n_{>n}$ . For this reason, the numerical values of  $p'_{DCG}$  given in table I correspond to theoretical bounds<sup>2</sup> achievable utilizing partial correlation of order up to  $n_{max} = 3$  under ideal experimental conditions for sparse random graphs. The influence of  $n_{max}$  from 0 to 3

TABLE I

NUMERICAL VALUES FOR  $p'_{DCG}$  FOR THE RESULTS SHOWN IN FIG. 4 TO 9 WITH  $p'_{DCG} = 1 - p_c + p_{>n}$ . THE SYMBOLS IN FRONT OF THE NETWORK PARAMETERS CORRESPOND TO THE CUMULATIVE DISTRIBUTIONS SHOWN IN FIG. 3.

(symbol) Network parameter	$p'_{DCG}$
+: $N = 20, q = 0.15$	0.81
x: $N = 20, q = 0.05$	0.89
◊: $N = 20, q = 0.075$	0.78
o: $N = 30, q = 0.10$	0.89
Δ: $N = 20, q = 0.10$	0.68
□: $N = 50, q = 0.03$	0.95

on the theoretical bound of the precision of inferable network structure is shown in Fig. 3. In general, the larger  $n_{max}$  the higher the theoretical bound. Interestingly, from these results

<sup>2</sup>Within the accuracy of our numerical simulations.

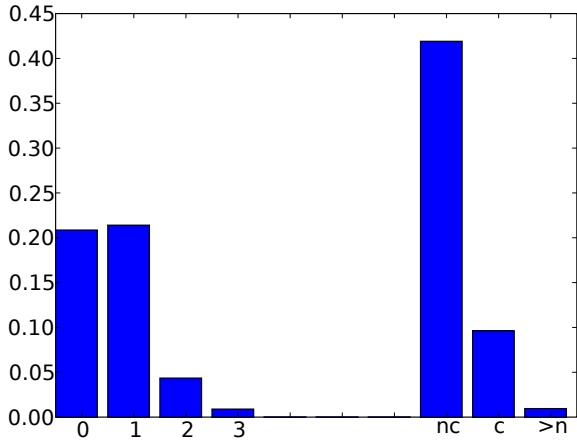


Fig. 4.  $N = 20, q = 0.05$  and  $z = 1$ .

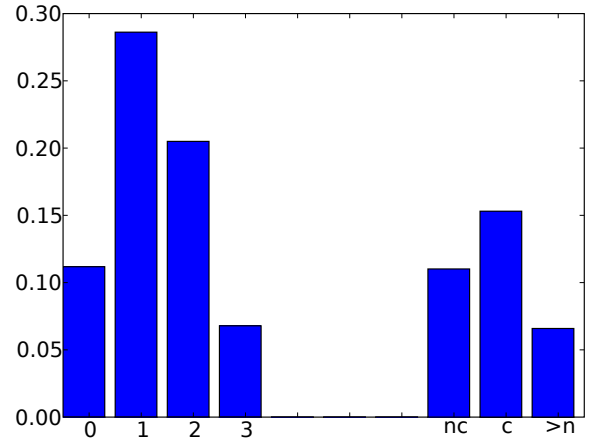


Fig. 7.  $N = 20, q = 0.10$  and  $z = 2$ .

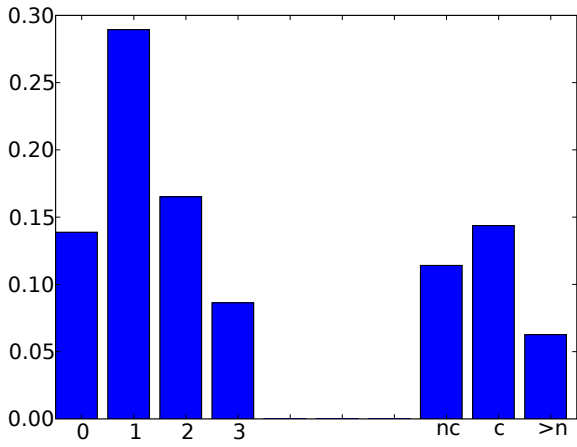


Fig. 5.  $N = 20, q = 0.075$  and  $z = 1.5$ .

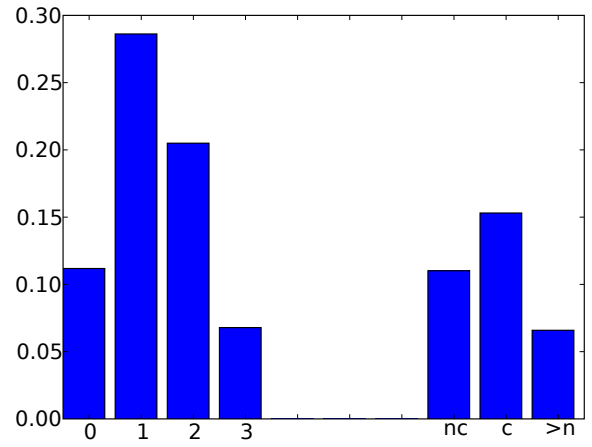


Fig. 8.  $N = 20, q = 0.15$  and  $z = 3$ .

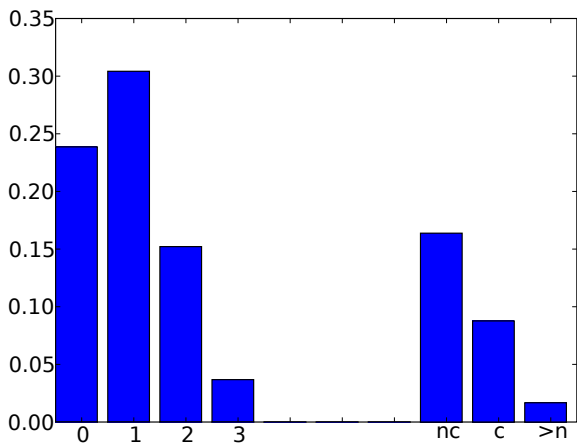


Fig. 6.  $N = 30, q = 0.10$  and  $z = 3$ . Lhs: results for d-separating sets of size 0 to  $n_{max} = 3$ . Rhs: values for nc = 'not connected', c = 'connected' and  $n_{<} =$  'not d-separable up to  $n_{max}$ '.

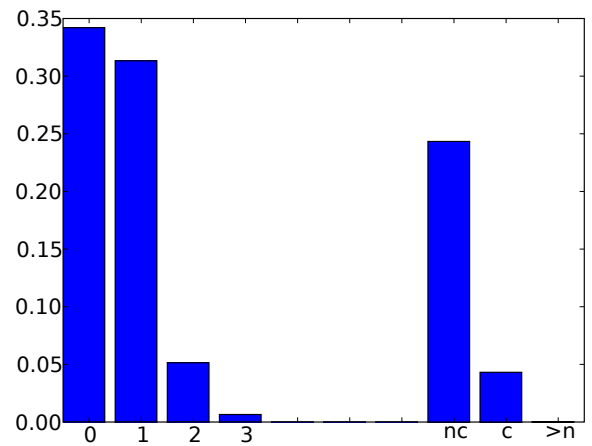


Fig. 9.  $N = 50, q = 0.03$  and  $z = 1.5$ . Lhs: results for d-separating sets of size 0 to  $n_{max} = 3$ . Rhs: values for nc = 'not connected', c = 'connected' and  $n_{<} =$  'not d-separable up to  $n_{max}$ '.

it is clear that even under ideal experimental conditions the networks structure can not be inferred perfectly if a method is applied solely based on partial correlations but with at least 70% or more if the underlying graph is sparse and random.

#### IV. CONCLUSIONS

Our numerical results demonstrate clearly, that considering d-separating sets only up to size  $n_{max} = 3$  results in an error less than 5% for a given network structure. From this one can conclude, that under ideal experimental conditions an algorithm utilizing only partial correlations up to order  $n_{max}$  to reconstruct the underlying network structure from ideal experimental data would make the same error. That means, the overall precision of the inferable network structure is given by  $p'_{DCG} = 1 - p_c + p_{>n}$ , because node pairs that are directly connected can not be distinguished by partial correlation from node pairs that are not d-separable with set sizes  $\leq n_{max}$ . Hence,  $p'_{DCG}$  is a theoretical bound achievable for inference algorithms solely based on partial correlations under ideal experimental conditions.

This confirms heuristic approaches used to reconstruct gene networks from low-order partial correlations [de la Fuente et al., 2004], [Magwene and Kim, 2004], [Wille and Bühlmann, 2006]. However, further studies are necessary to demonstrate that these results hold also for different network classes besides sparse random networks, e.g., scale-free, small-world or hierarchical networks whose structure is more close to the structure of biological gene networks as found by, e.g., [Barabasi and Oltvai, 2004], [Basso et al., 2005], [Holme et al., 2003], [Ravasz et al., 2002] and also by [van Noort et al., 2004].

*Acknowledgments:* We would like to thank Jie Chen, Alberto de la Fuente, Earl F. Glynn, Bill Shipley and Korbinian Strimmer for fruitful discussions.

*References:*

#### REFERENCES

- [Barabasi and Oltvai, 2004] Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews*, 5:101–113.
- [Basso et al., 2005] Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human b cells. *Nature Genetics*, 37(4):382–390.
- [de la Fuente et al., 2004] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574.
- [Erdős and Rényi, 1959] Erdős, P. and Rényi, A. (1959). On random graphs. *Publitiones Mathematicae*, 6:290–297.
- [Holme et al., 2003] Holme, P., Huss, M., and Jeong, H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4):532–538.
- [Jeong et al., 2001] Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411:41–42.
- [Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407:651–654.
- [Magwene and Kim, 2004] Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, 5(12):R100.

- [Newman, 2003] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.
- [Pearl, 1988] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufmann.
- [Pearl, 2000] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge.
- [Ravasz et al., 2002] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.
- [Solomonoff and Rapoport, 1951] Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biophysics*, 13:107–117.
- [Spirtes and Glymour, 1991] Spirtes, P. and Glymour, C. (1991). A algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9:62–72.
- [Spirtes et al., 2000] Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.
- [Spirtes et al., 1998] Spirtes, P., Richardson, T., Meek, C., Scheines, R., and Glymour, C. (1998). Using path diagrams as a structural equation modelling tool. *Sociological Methods and Research*, 27:182–225.
- [van Noort et al., 2004] van Noort, V., Snel, B., and Huymen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO reports*, 5(3):280–284.
- [Verma and Pearl, 1988] Verma, T. and Pearl, J. (1988). Causal networks: semantics and expressiveness. In *Proceedings of the 4th workshop on uncertainty in artificial intelligence*, pages 352–359. Mountain View CA.
- [Verma and Pearl, 1991] Verma, T. and Pearl, J. (1991). Equivalence and synthesis of causal models. In *Proceedings of the 6th workshop on uncertainty in artificial intelligence*, pages 220–227. Cambridge, MA.
- [Wille and Bühlmann, 2006] Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical Applications in Genetics and Molecular Biology*, 4(1):32.