Learning Belief Network Structure From Data under Causal Insufficiency

Mieczysław A. Kłopotek

Institute of Computer Science, Polish Academy of Sciences

e-mail: klopotek@ipipan.waw.pl

1 Introduction

Various expert systems, dealing with uncertain data and knowledge, possess knowledge representation in terms of a belief network (e.g. knowledge base of the MUNIM system [1] , ALARM network [2] etc.). A number of efficient algorithms for propagation of uncertainty within belief networks and their derivatives have been developed, compare e.g. [9], [11], [12].

Belief networks, causal networks, or influence diagrams, or (in Polish) cause-effect networks are terms frequently used interchangeably. They are quite popular for expressing causal relations under multiple variable setting both for deterministic and non-deterministic (e.g. stochastic) relationships in various domains: statistics, philosophy, artificial intelligence [3], [14]. Though a belief network (a representation of the joint probability distribution, see [3]) and a causal network (a representation of causal relationships [14]) are intended to mean different things, they are closely related. Both assume an underlying dag (directed acyclic graph) structure of relations among variables and if Markov condition and faithfulness condition [15] are met, then a causal network is in fact a belief network. The difference comes to appearance when we recover belief network and causal network structure from data. A dag of a belief network is satisfactory if the generated probability distribution fits the data, may be some sort of minimality is required. A causal network structure may be impossible to recover completely from data as not all directions of causal links may be uniquely determined [15]. Fortunately, if we deal with causally sufficient sets of variables (that is whenever significant influence variables are not omitted from observation), then there exists the possibility to identify the family of belief networks a causal network belongs to [16].

Regrettably, to our knowledge, a similar result is not directly known for causally insufficient sets of variables (that is when significant influence variables are hidden) - "Statistical indistinguishability is less well understood when graphs can contain variables representing unmeasured common causes" ([15], p. 88). Latent (hidden) variable identification has been investigated intensely both for belief networks (e.g. [8], [6], [7], [2]) and causal networks ([10], [14], [15], [4], [5]), as well as in traditional statistics (see [13] for a comparative study of LISREL and EQS techniques). The algorithm of [2] recovers the most probable location of a hidden variable. Whereas the CI algorithm of [15] recovers exact locations of common causes, but clearly not all of them. In fact, the CI algorithm does not provide a dag, but rather a graph with edges fully (unidirected or bidirected) or partially oriented, or totally non-oriented with additional constraints for edge directions at other edges. Partially or non-oriented edges may prove to be either directed or bidirected edges.

The big open question is whether or not the bidirectional edges (that is indications of a common cause) are the only ones necessary to develop a belief

network out of the product of CI, or must there be some other hidden variables added (e.g. by guessing). This paper is devoted to settling this question.

2 Causal Inference Algorithm

Below we remind the Causal Inference (CI) algorithm of Spirtes, Glymour and Scheines [15] together with some basic notation used therein.

Essentially, the CI algorithm recovers partially the structure of an including path graph. Given a directed acyclic graph G with the set of hidden nodes V_h and visible nodes V_s representing a causal network CN, an including path between nodes A and B belonging to V_s is a path in the graph G such that the only visible nodes (except for A and B) on the path are those where edges of the path meet head-to-head and there exists a directed path in G from such a node to either A or B. An including path graph for G is such a graph over V_s in which if nodes A and B are connected by an including path in G ingoing into A and B, then A and B are connected by a bidirectional edge $A <-> B$. Otherwise if they are connected by an including path in G outgoing from A and ingoing into B then A and B are connected by an unidirectional edge $A-> B$.

A partially oriented including path graph contains the following types of edges unidirectional: $A- > B$, bidirectional $A < - > B$, partially oriented $Ao- > B$ and non-oriented $Ao - oB$, as well as some local constraint information $A * -\underline{*B*} - *C$meaning that edges between A and B and between B and C cannot meet head to head at B. (Subsequently an asterisk (*) means any orientation of an edge end: e.g. $A * - > B$ means either $A- > B$ or $Ao- > B$ or $A < - > B$).

In a partially oriented including graph π:

(i) A is a parent of B if and only if $A- > B$ in π.

(ii) B is a collider along the path $< A, B, C >$ if and only if $A * - > B < - * C$ in π.

(iii) An edge between B and A is into A iff $A < - * B$ in π

(iv) An edge between B and A is out of A iff $A- > B$ in π.

(v) A is d-separated from B given set S iff A and B are conditionally independent given S.

(vi) A and B are d-connected given C iff there exists no such S containing C such that A and B are conditionally independent given S.

(vii) In a partially oriented including path graph π', U is a definite discriminating path for B if and only if U is an undirected path between X and Y containing B, $B \neq X, B \neq Y$, every vertex on U except for B and the endpoints is a collider or a definite non-collider on U and:

(a) if V and V" are adjacent on U, and V" is between V and B on U, then $V * - > V$" on U,

(b) if V is between X and B on U and V is a collider on U, then $V- > Y$ in π, else $V < - * Y$ on π

(c) if V is between Y and B on U and V is a collider on U, then $V- > X$ in π, else $V < - * X$ on π

(d) X and Y are not adjacent in π.

viii) Directed path U: from X to Y: if V is adjacent to X on U then $X- > V$ in π, if V is adjacent to Y on V, then $V- > Y$, if V and V" are adjacent on U and V is between X and V" on U, then $V- > V$" in π.

The Causal Inference (CI) Algorithm:

Input: Empirical joint probability distribution

Output: partial including path graph π.

A) Form the complete undirected graph Q on the vertex set V.

B) if A and B are d-separated given any subset S of V, remove the edge between A and B, and record S in Sepset(A,B) and Sepset(B,A).

C) Let F be the graph resulting from step B). Orient each edge o-o. For each triple of vertices A,B,C such that the pair A,B and the pair B,C are each adjacent in F, but the pair A,C are not adjacent in F, orient (C) A*-*B*-*C as $A * - > B < - * C$ if and only if B is not in Sepset(A,C), and orient A*-*B*-*C as $A * - \underline{* B *} - * C$ if and only if B is in Sepset(A,C).

D) Repeat

if there is a directed path from A to B, and an edge A*-*B, orient (D_p) A*-*B as $A * - > B$,

else if B is a collider along $< A, B, C >$ in π, B is adjacent to D, and A and C are not d-connected given D, then orient (D_s) $B * - * D$ as $B < - * D$,

else if U is a definite discriminating path between A and B for M in π and P and R are adjacent to M on U, and P-M-R is a triangle, then
if M is in Sepset(A,B) then M is marked as non-collider on subpath $P * - \underline{* M *} - R$
else $P * - * AM * - * R$ is oriented (D_d) as $P * - > M < - * R$,

else if $P * - \underline{> M *} - * R$ then orient (D_c) as $P * - > M - > R$.

until no more edges can be oriented.

End of CI

3 From CI Output to Belief Network

Let us imagine that we have obtained a partial including path graph from CI, and we want to find a Belief Network representing the joint probability distribution out of it. Let us consider the following algorithm:

CI-to-BN Algorithm

Input: Result of the CI algorithm (a partial including path graph)

Output: A belief network

A) Accept unidirectional and bidirectional edges obtained from CI.

B) Orient every edge $Ao- > B$ as $A- > B$.

C) Orient edges of type $Ao - oB$ either as $A < -B$ or $A- > B$ so as not to violate $P * -\underline{*M*} - *R$ constraints.

End of CI-to-BN

We claim that:

THEOREM 1 *(i) By the CI-to-BN algorithm, a belief network can always be obtained.*
(ii) The obtained belief network keeps all the dependencies and independencies of the original underlying including path graph.

The rest of this section provides a sketchy proof of the above theorem.

First, we shall notice, that hiding variables - while keeping the connections of original variables - introduces into an including path graph two kinds of new edges: (1) those stemming from a direct connection: unidirectional, keeping the acyclicity of the graph, (ii) common cause connections, introducing bidirectionality, Let us notice, that the dag of the original network (the intrinsic one) introduced a partial ordering of the (observed and unobserved) nodes. If a new (uni)directed edge is introduced by variable hiding, then only if there existed a directed path between the nodes. Therefore, the original partial order is not violated by them. On the other hand, a newly introduced bidirectional edge indicates that there exists a hidden node such that there exist a directed path from this node to each of the now connected nodes in the original graph. So it is justified to replace the introduced bidirectional edge by a parentless common cause and in this way also the partial ordering of the original graph is not violated. Hence the obtained including path graph is in fact a dag. Also, all the dependencies and independences represented by this graph would be really those which are present in the original graph under the assumption of variable hiding. There remains only one cosmetic question to be settled: What happens if a new bidirectional edge is introduced between nodes which were already connected within the original dag. Should we keep the information that there was an edge between the nodes in the original graph between them or can we drop it without affecting the dependence/independence information ? It happens that we can drop it because the procedure introduces necessary additional edges around this connection (vital for incoming directed edges) in order to keep all the paths considered by d-separation intact. Hence it is legitimate to draw a dag (called below the full hiding dag) consisting only of observable nodes with uni- and bi-directional edges the latter ones indicating that the nodes are not connected but have a hidden common cause.

Now let us consider the following question: Let us consider a mixed graph (MG) having all the unoriented edges of the original including path graph (here called FHD - Full Hiding Dag) and the CI partially oriented including path graph (here called CIG - CI graph). Let all the unidirected edges of FHD be unidirected the same way in MG, let all the edges bidirected by CI he bidirected in MG. Let all partially directed edges of CI be unidirected in MG. Last not least, let all bidirectional FHD edges not oriented at all by Ci be left unoriented in MG. Now if there were no cycles in MG, then also the claim (i) of the above theorem would be valid. The proof of acyclicity of MG is not difficult, but laborious. An overview can be made in terms of Figures, from Fig.2 to Fig.11. In this series of Figures it is demonstrated that no three edges of MG can form a cycle.
Within the CI algorithm, only orientation steps denoted as D_p, D_s and C can give rise to a partial orientation of a FHD-bidirectional edge within CIG. (Step D_d immediately creates a bidirectional edge in CIG out of it). The Figures summarize the proof as follows: e.g. in Fig.4 (a) overviews the general situation in the FHD when one edge (BA) is bidirectional and AC and CB are unidirectional and the edge DA caused in step (C) of CI orientation of BA and DA towards A. (b) and (c) follow the case when the edges DA and AC are not bridged: as a result the edge BA is made biorriented in CIG. On the other hand, (d) and (f) deal with the case when DA and AC are bridged: then also BA grows bidirectional in CIG. So. Fig.3-4 show that in case of one FHD-bidirectional and two FHD-unidirectional edges no cycle in MG is possible. Fig.5-9 demonstrate the same for two FHD-bi- and one FHDunidirectional edges, demonstrating, that both bidirectional edges will be made bidirectional in MG if there were any risk of cyclicity during the CI-algorithm. Fig.10-11 are concened with potential cycles consisting of three FHD-bidirectional edges
Within a longer trail of edges it is immediately visible, that there must exist at least one pair of neighboring edges (one of them bidirectional in FHD) which are "bridged" that is their ends not neighboring on the path are neighbors in the graph. Hence we have here a triangle which - due to facts proven earlier - cannot form a cycle in MG and at least two edges and at least one FHD-bidirectional are oriented correctly (that is as in FHD), hence cannot participate also in a larger cycle. (This is clear if the "third" edge has also been

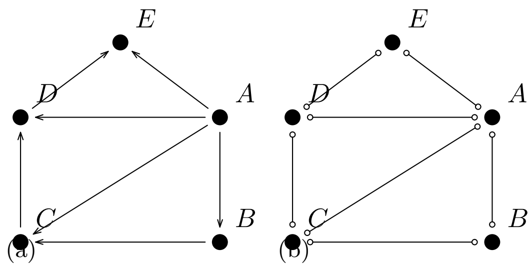

Figure 1: A result of running CI algorithm: (a) the original dag (b) the dag recovered

oriented by C, D_s or D_p step. But we can easily check, that if it has been oriented by the D_d step, then the other edges will be oriented prohibiting a long-run cycle.) This completes the proof of claim (i).

As claim (ii) is concerned, we shall first notice that a situation like that of Fig.12 cannot happen in an including path graph, that is it is never possible, that along a path $AB_1...B_nC$ with head to head meetings at B_i one edge outgoing from each B_i points at A and there is some j such that an edge B_jC is outgoing from B_j. Now, when orienting edges according to CI-to-BN algorithm, we can make two types of errors:(a) introduce a path which is not active (in terminology of [3]) in BN, but is actually active in FHD, and (b) introduce a path which is active (in terminology of [3]) in BN, but is actually not active (blocked) in FHD. In case (a), we may have the structure of such a path as $..., D, B_n, ..., B_1, A, C_1, ..., C_m, E, ...$ in Fig. 13, with node A set in BN erroneously active (to the left) or passive (to the right). Let us assume that this is the shortest active path between the nodes of interest that is no subset of nodes on the erroneously active path can form also an active path. Then in Fig.13.a) and .b) there exists no unioriented edge $D- > B_i$ nor $E- > C_j$, nor bidirectional edge $B_i < - > C_j$ nor $D < - > B_j$ nor $E < - > C_j$ nor $D < - > A$ nor $E < - > A$,for any i,j. And additionally in Fig.13.a) there exists no edge $D- > A$ nor $E- > A$. In Fig.13.b) there exists no edge $D < -A$ nor $E < -A$. But it can then be demonstrated, that the CI orients correctly nodes from D to B_1 and from E to C_1, and then a definite discriminating path for A emerges, and the edges at A are oriented correctly, hence it is denied that an error may occur at A.

As error (b) is concerned, we can proceed in an analogous way, also assuming that we have to do with the shortest erroneously passive path.

This would then complete the proof of the Theorem.

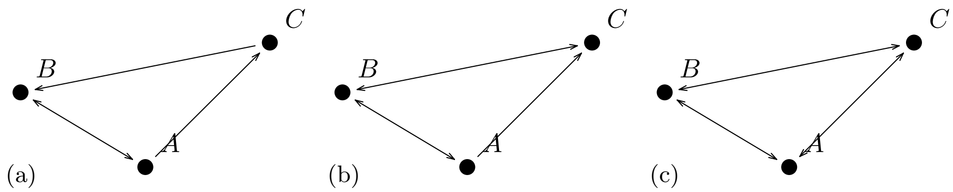

Figure 2: Δ - a triangle in the original including path graph.(a)- $\Delta 1$ - only one intrinsic bidirectional edge(b)- $\Delta 2$ - only two intrinsic bidirectional edges(c)- $\Delta 3$ - three bidirected edges

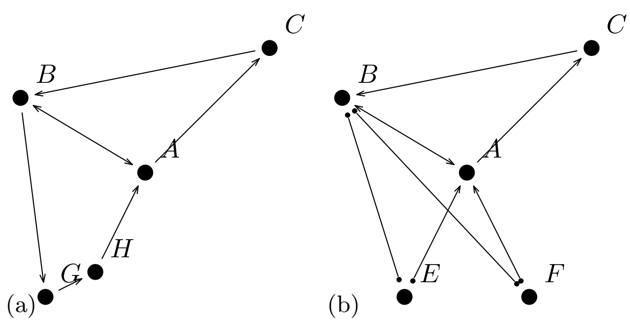

Figure 3: (a)- $\Delta 1 D_p$ - orientation due to a directed path - impossible as introducing a cycle(b)- $\Delta 1 D_s$ - orientation due to a separation - impossible as oriented path $A- > C- > B$ renders E,F dependent given B (denying defining condition of (D_s) for arrow $B * - > A$)

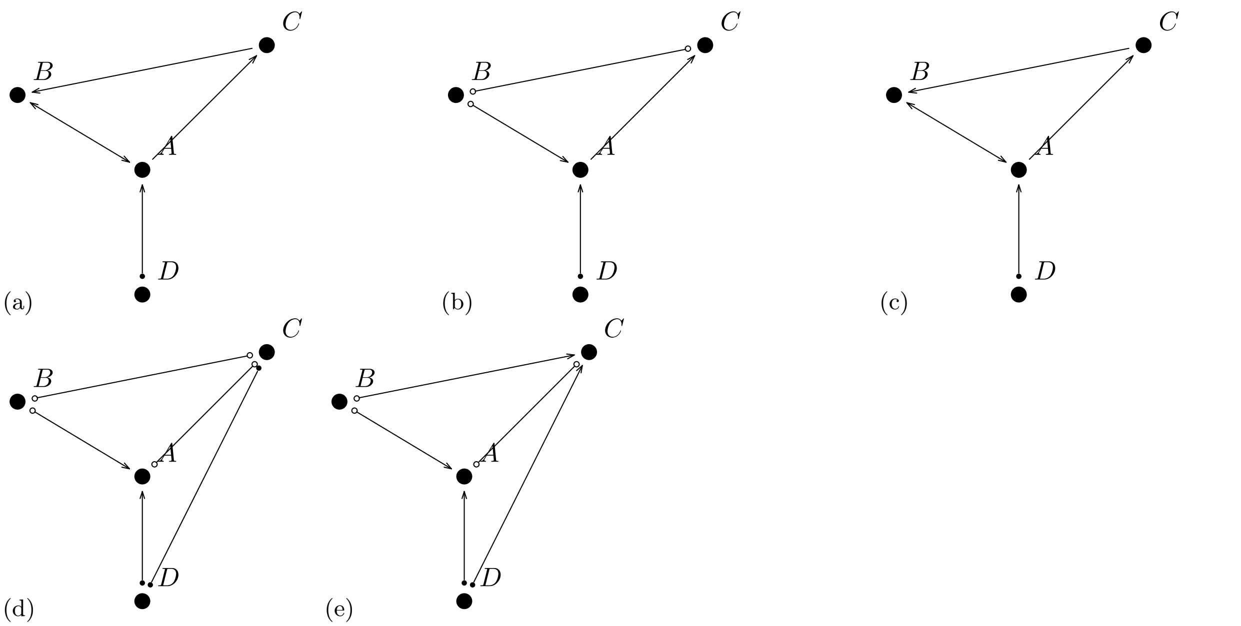

Figure 4: (a)- $\Delta 1C$ - orientation due to unbridged head to head(b)- $\Delta 1C(1)$ - case 1 (CIA-HD) - due to (C) on path DAC(c)- $\Delta 1C(1)b$ - hence due to discriminating path DACB for C and DABC for B - O.K.(d)- $\Delta 1C(2)$ - case 2 (CIA-HD) - when DAC is bridged(e)- $\Delta 1C(2)b$ - hence due to (C) on DCBand due to A being a collider along DAB and the edge $A->C$- but this is impossible as it denies the intrinsic structure for edge $C->B$

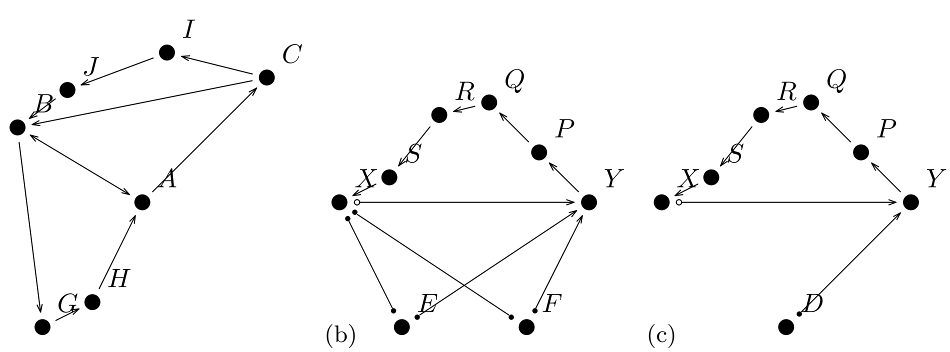

Figure 5: (a)- $\Delta 2D_pD_p$ - impossible -due to a cycle(b)- $\Delta 2D_pD_s$ - impossible -due denying orientation capability of $X*->Y$(c)- $\Delta 2D_pC$ - impossible -due to denying non-connection of X and D

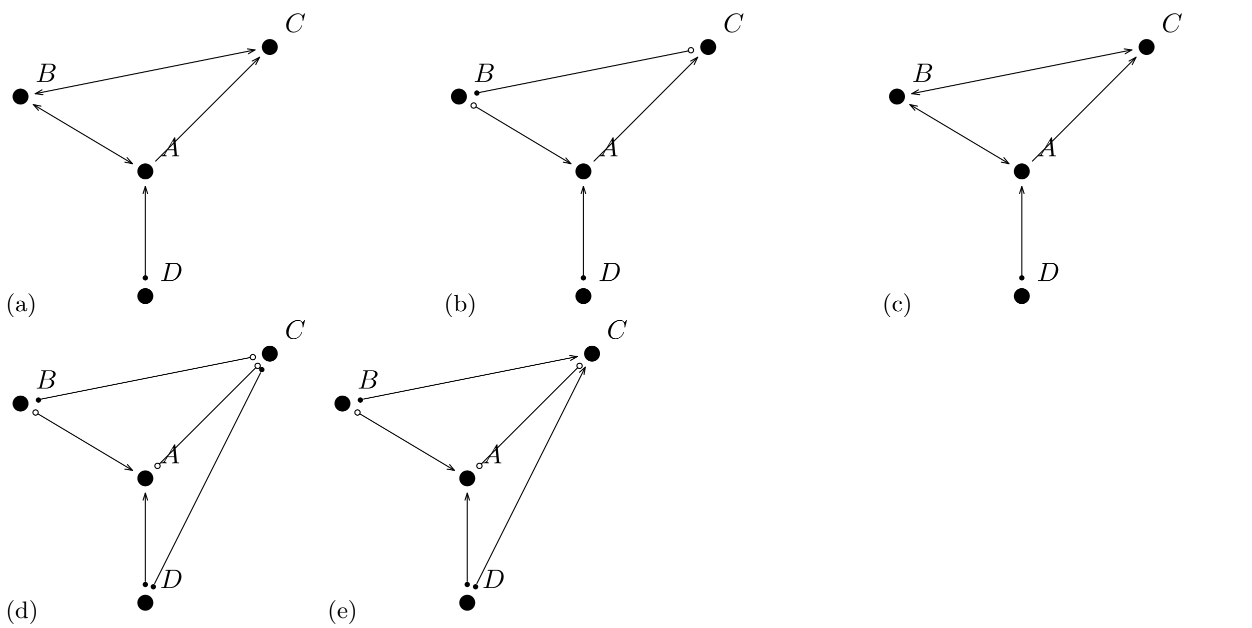

Figure 6: (a)- $\Delta 2C@A$ - case (C)-orientation applied at A(b)- $\Delta 2C@A(1)$ - case DAC not bridged(c)- $\Delta 2C@A(1)b$ - hence, due to discriminating path DACB for C and DABC for B - O.K.(d)- $\Delta 1C@A(2)$ - case 2 (CIA-HD) - when DAC is bridged(e)- $\Delta 1C@A(2)b$ - hence due to (C) on DCBand due to A being a collider along DAB and the edge $A- > C$- O.K.

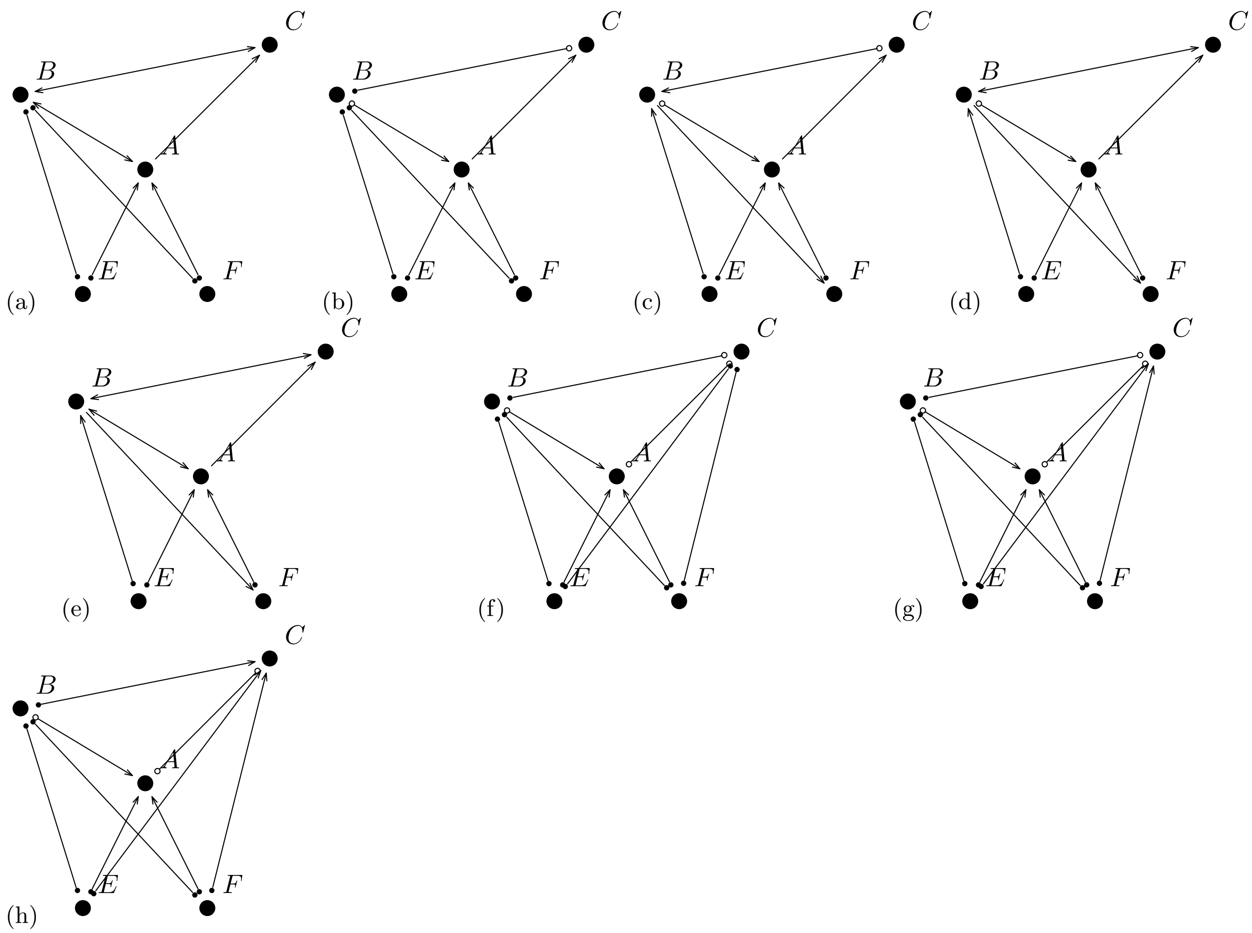

Figure 7: (a)- $\Delta 2 D_s@A$ - case (D_s)-orientation applied at A(b)- $\Delta 2 D_s@A(1)$ - case EAC (or by analogy: FAC) not bridged(c)- $\Delta 2 D_s@A(1)b$ - hence due to non-bridged path EBC and EBF (d)- $\Delta 2 D_s@A(1)c$ - hence due to definite discriminating path EACB for C(e)- $\Delta 2 D_s@A(1)d$ - hence due to definite discriminating path EABC for B - O.K.(f)- $\Delta 2 D_s@A(2)$ - case both EAC and FAC bridged(g)- $\Delta 2 D_s@A(2)b$ - hence due to unbridged ECF(h)- $\Delta 2 D_s@A(2)c$ - hence due to d-separability of E,F given B - O.K.

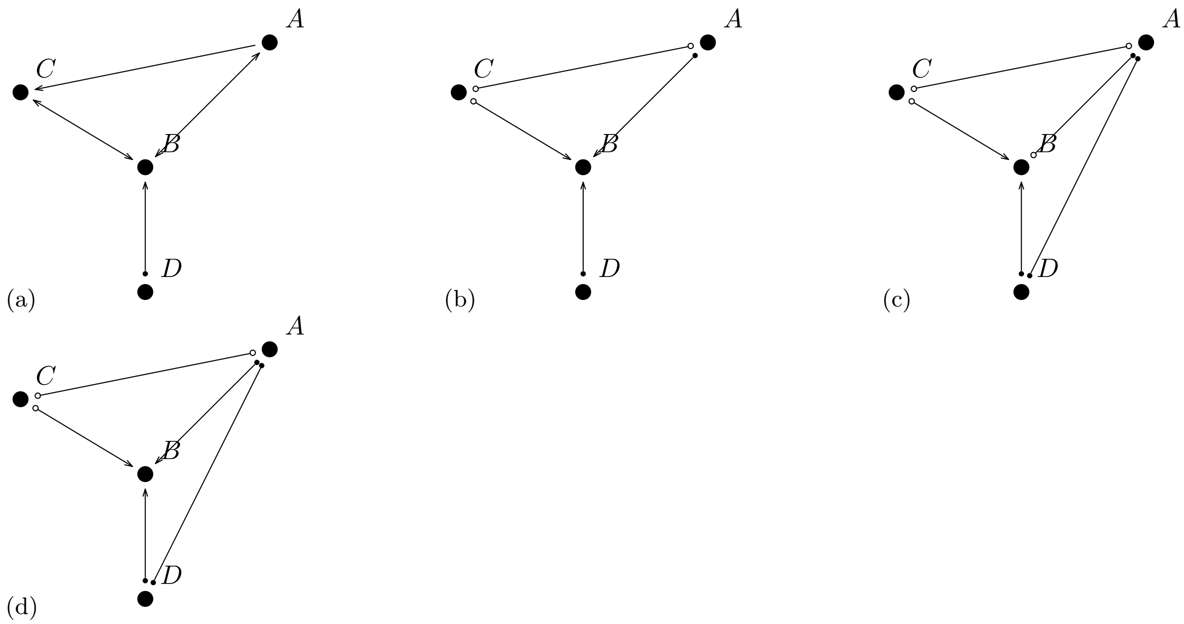

Figure 8: (a)- $\Delta 2C@B$ - case (C)-orientation applied at B(b)- $\Delta 2C@B(1)$ - case DAC not bridged - O.K.(c)- $\Delta 2C@B(2)$ - case DAC bridged(d)- $\Delta 2C@B(2)b$ - due to (D_s at A for collider $C*->B<-*A$)- O.K.

4 Summary and Outlook

Within this paper an algorithm of recovery of belief network structure from data has been presented and its correctness demonstrated. It relies essentially on exploitation of the result of the known CI algorithm of Spirtes, Glymour and Scheines [15]. The edges of partial including path graph, not oriented by CI, are oriented to form a directed acyclic graph. The contribution of this paper is to show that such an orientation of edges always exists without necessity of adding auxiliary hidden variables, and that this dag captures all dependencies and independencies of the intrinsic underlying including path graph.

The CI-to-BN algorithm will suffer from the very same shortcomings as the CI algorithm, that is it is tractable only for a small number of edges. It will be interesting task to examine the possibility of such an adaptation of the Fast CI algorithm [15]. It may not be trivial as the product of CI differs from that of FCI [15]. Another path of research would be to elaborate of version of CI-to-BN assuming only with a restricted number of variables participating in a d-separation, which would also bind the exponential explosion of numerical complexity.

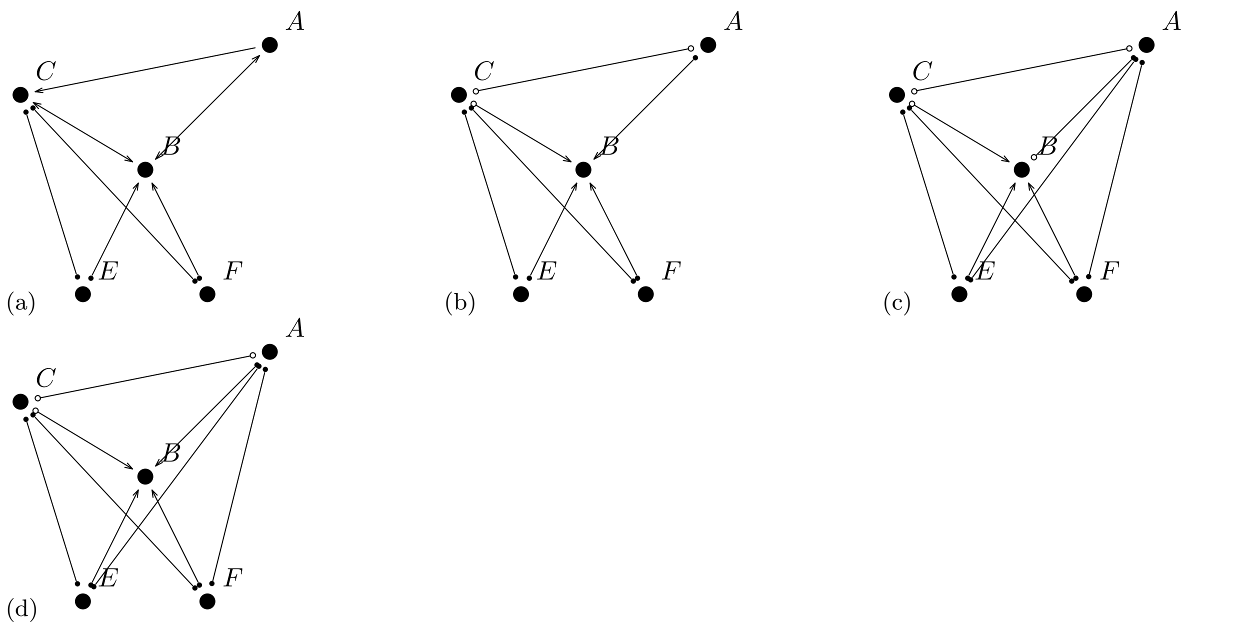

Figure 9: (a)- $\Delta 2D_s@B$ - case (D_s)-orientation applied at B(b)- $\Delta 2D_s@B(1)$ - EBA (or by analogy: FBA) not bridged- O.K.(c)- $\Delta 2D_s@B(2)$ - EBA and FBA bridged(d)- $\Delta 2D_s@B(2)b$ - due to (D_s at A for collider $E* -> B <- *F$O.K.

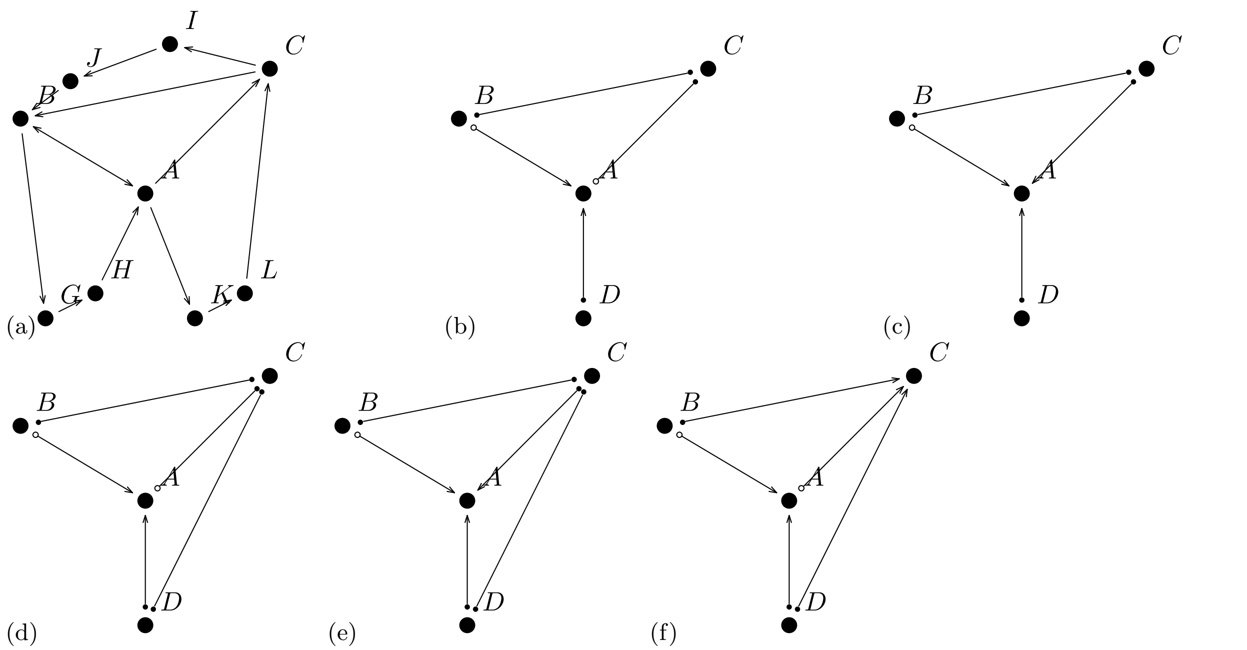

Figure 10: (a)- $\Delta 3 D_p D_p D_p$ - impossible -due to a cycle(b)- $\Delta 3C$ - full symmetry in all nodes A,B,C(c)- $\Delta 3C(1)$ - DAC unbridged - O.K.(d)- $\Delta 3C(2)$ - DAC bridged(e)- $\Delta 3C(2)b(1)$ - if (D_s) at node C for DAB applicable- O.K.(f)- $\Delta 3C(2)b(2)$ - if (D_s) at node C for DAB not applicable - O.K. (D_p for $C* -> B$ impossible as non-connection of D and B would be denied.

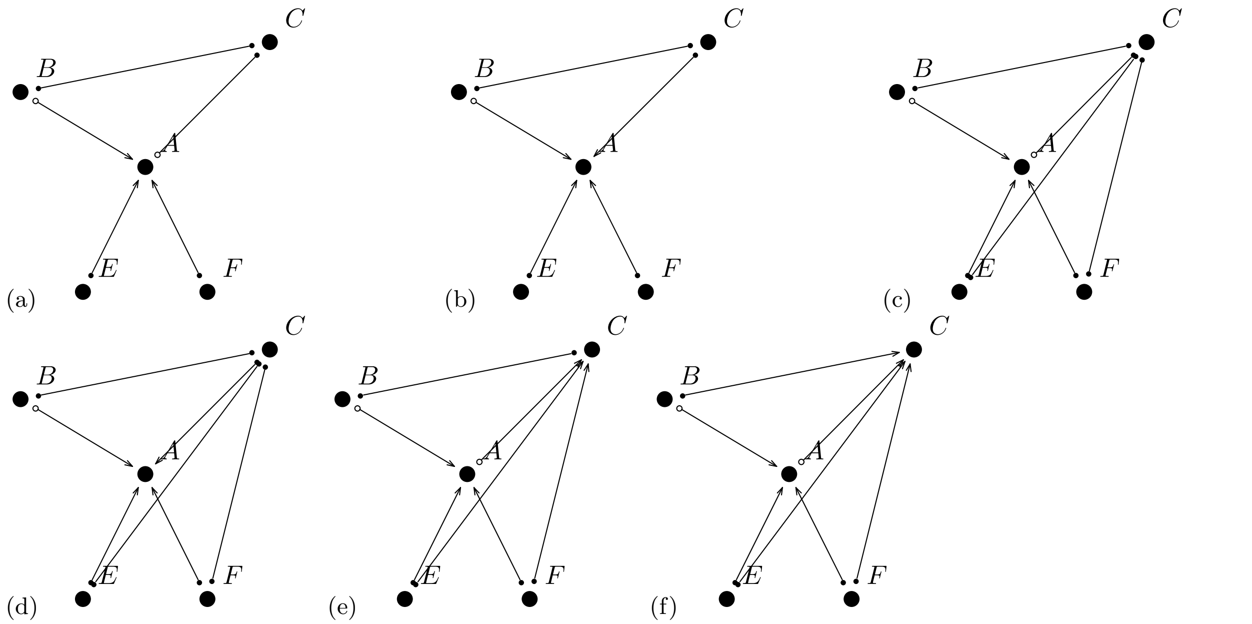

Figure 11: (a)- $\Delta 3D_s$ - full symmetry in all nodes A,B,C(b)- $\Delta 3D_s(1)$ - EAC unbridged - O.K.(c)- $\Delta 3D_s(2)$ - FAC and EAC bridged(d)- $\Delta 3D_s(2)b(1)$ - if (D_s) at node C for EAF applicable- O.K.(e)- $\Delta 3D_s(2)b(2)$ - if (D_s) at node C for EAF not applicable,(f)- $\Delta 3D_s(2)b(2)b$ - then (D_s at node B for ECF applicable, (D_p for $C*->B$ would then not be appropriate)

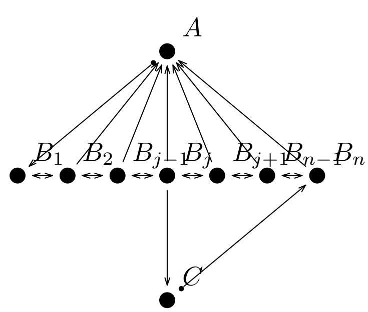

Figure 12: Impossible because B_j cannot participate in d-separation of A and C

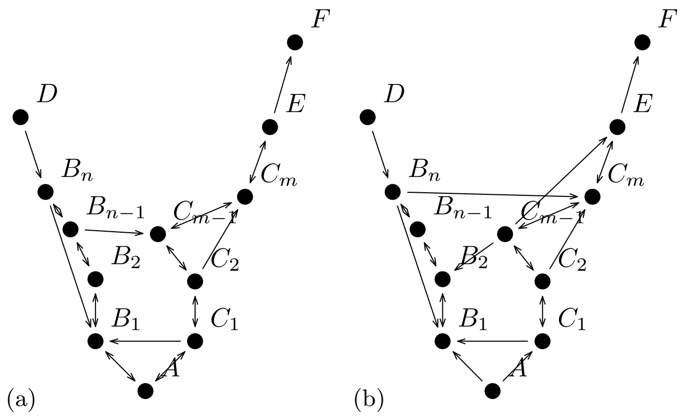

Figure 13: hdags possibly causing node A to be erroneously considered as (a) non-collider (b) a collider.

References

[1] Andreassen S., Woldbye M., Falck B., Andersen S.K.: MUNIN - a causal probabilistic network for interpretation of electromyographic findings, *Proc. of the Tenth International Joint Conference on AI*, 366-372, 1987.

[2] Cooper G.F., Herskovits E.: *A Bayesian method for the induction of probabilistic networks from data*, Machine Learning **9** (1992), 309-347.

[3] Geiger D., Verma T, Pearl J.: *d-Separation: From theorems to algorithms*, M.Henrion, R.D.Shachter, L.N.Kamal, J.F.Lemmer (Eds): Uncertainty in Artificial Intelligence 5, Elsevier Science Publishers B.V. (North-Holland), 1990, pp. 139-148.

[4] Glymour C., Scheines R., Spirtes P., Kelley K.: *Discovering causal structure*, New York, Academic Press, 1987.

[5] Glymour C., Spirtes P.: Latent variables, causal models and overidentifying constraints, *Journal of Econometrics*, 39(1988), 175-198.

[6] Golmard J.L., Mallet A.: Learning probabilities in causal trees from incomplete databases, *Proc. of the IJCAI Workshop on Knowledge Discovery in Databases (Detroit M.I.)*, (1989), 117-126

[7] Liu L., Wilkins D.C., Ying X., Bian Z.: Minimum error tree decomposition, *Proc. of the Conference on Uncertainty in Artificial Intelligence*, Cambridge M.A.,(1990) 54-62

[8] Pearl J.: Fusion, propagation and structuring in belief networks, *Artificial Intelligence*, 29 (1986), 241-288.

[9] Pearl J.: *Probabilistic Reasoning in Intelligent Systems:Networks of Plausible Inference*, Morgan Kaufmann, San Mateo CA, 1988

[10] Pearl J., Verma T.: A theory of inferred causation,, [in:] Principles of Knowledge Representation and Reasoning, Proc. of the Second International Conference, Cambridge, Massachusetts, April 22-25, 1991, Allen J., Fikes R., Sandewell E. Eds,, San Mateo CA:Morgen Kaufmann, 441-452.

[11] Shachter R.D.: *Evidence absorption and propagation through evidence reversals* , in: M. Henrion, B.D. Shachter, L.N. Kanal, J.F. Lemmer: Uncertainty in Artificial Intelligence 5, Elsevier Science Publishers B.V (North-Holland) (1990), 173- 190

[12] Shenoy P.P., Shafer G.: *Axioms for probability and belief-function propagation*, in: R.D. Shachter, T.S. Levit, L.N. Kanal, J.F. Lemmer eds: Uncertainty in Artificial Intelligence 4, (Elsevier Science Publishers B.V. (North Holland), 1990).

[13] Spirtes P., Glymour C., Scheines R.: Simulation studies of the reliability of computer-aided model specification using TETRAD II, EQS and LISREL programs, *Sociological Methods and Research*, 19 (1990), 3-66

[14] Spirtes P., Glymour C., Scheines R.: *Causality from probability*, in: G.McKee ed.: Evolving knowledge in natural and artificial intelligence (London: Pitman, 1990), 181-199.

[15] Spirtes P., Glymour C., Scheines R.: Causation, Prediction and Search, Lecture Notes in Statistics 81, Springer-Verlag, 1993.

[16] Verma T., Pearl J.: Equivalence and synthesis of causal models, Proc. 6th Conference on Uncertainty in AI, 220-227, 1990.