Characterization of Inclusion Neighbourhood in Terms of the Essential Graph: Upper Neighbours*

Milan Studený**

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
Pod vodárenskou věží 4, 18208 Prague, Czech Republic
studeny@utia.cas.cz

Abstract. The problem of efficient characterization of inclusion neighbourhood is crucial for some methods of learning (equivalence classes of) Bayesian networks. In this paper, neighbouring equivalence classes of a given equivalence class of Bayesian networks are characterized efficiently by means of the respective essential graph. The characterization reveals hidded internal structure of the inclusion neighbourhood. More exactly, upper neighbours, that is, those neighbouring equivalence classes which describe more independencies, are completely characterized here. First, every upper neighbour is characterized by a pair $([a,b],C)$ where $[a,b]$ is an edge in the essential graph and $C \subseteq N \setminus \{a,b\}$ a disjoint set of nodes. Second, if $[a,b]$ is fixed, the class of sets C which characterize the respective neighbours is a *tuft* of sets determined by its least set and the list of its maximal sets. These sets can be read directly from the essential graph. An analogous characterization of lower neighbours, which is more complex, is mentioned.

1 Motivation

1.1 Learning Bayesian Networks

Several approaches to learning Bayesian networks use the method of maximization of a *quality criterion*, named also 'quality measure' [3] and 'score metric' [4]. Quality criterion is a function, designed by a statistician, which ascribes to data and a network a real number which 'evaluates' how the statistical model determined by the network is suitable to explain the occurence of data. Since the actual aim of the learning procedure is to get a statistical model (defined by a network) reasonable quality criteria do not distinguish between equivalent Bayesian networks, that is, between networks which define the same statistical model. Therefore, from operational point of view, the goal is to learn an equivalence class of Bayesian networks, that is, a class of acyclic directed graphs (over a fixed set of nodes N).

* This is an extended version with the Appendix which contains the proofs.

** This research has been supported by the grant GA ČR n. 201/01/1482.

As direct maximization of a quality criterion is typically infeasible the *method of local search* is often used. The main idea of this approach is that suitable concept of neighbourhood is introduced for acyclic directed graphs over N. The point is that the change in the value of a (reasonable) quality criterion is easy to compute for neighbouring graphs. Thus, instead of global maximization of a quality criterion one searches for a local maximum of the criterion with respect to the considered neighbourhood structure and this task is usually computationally feasible. Typical neighbourhood structures used in practice are defined by means of simple graphical operations with considered graphs - for details see [5, 7].

The algorithms of this kind can also be classified according to the method of representation of equivalence classes of networks. In some algorithms, an equivalence class is represented by any of its members which may, however, result in computational complications. In other algorithms, a special representative of each equivalence class is used. The most popular representative of an equivalence class of Bayesian networks is the *essential graph* which is a certain chain graph describing some common features of acyclic directed graphs from the class. The term 'essential graph' was proposed by Andersson, Madigan and Perlman [1]; altenative names 'completed pattern' [18], 'maximally oriented graph for a pattern' [9] and 'completed pdag' [5] have also appeared in the literature.

1.2 Inclusion Neighbourhood

There exists a neighbourhood structure (for equivalence classes of Bayesian networks) which has a good theoretical basis. The inclusion of statistical models defined by the networks, which corresponds to the inclusion of conditional independence structures defined by the network, induces a natural *inclusion ordering* on the collection of equivalence classes. This ordering induces a neighbourhood concept then. More specifically, two different types of neighbouring equivalence classes are assigned to every equivalence class of networks: the *upper* neighbours and *lower* neighbours. Thus, the *inclusion neighbourhood*, sometimes also named 'inclusion boundary neighbourhood' [7], consists of these two parts. There are also some practical reasons for using the inclusion neighbourhood - for details see [4]. Note that Chickering [5] has recently confirmed Meek's conjecture [10] about transformational characterization of the inclusion ordering. A consequence of this result is a graphical description of the inclusion neighbourhood in terms of the collection of members of the considered equivalence class (see Section 2.4).

The topic of this contribution is to characterize the inclusion neighbourhood of a given equivalence class of Bayesian networks in terms of the respective essential graph in such a way that it can be used efficiently in a method of local search for maximization of a quality criterion. Two recent papers were devoted to this problem, but, in author's view, none of them brought a satisfactory solution to the problem.

Chickering, in Section 5 of [5], gave a method which is able to generate tentatively all neighbouring equivalence classes (of a given equivalence class described by the respective essential graph). More specifically, two (composite) graphical operations applicable to an essential graph and respective legality tests which are able to decide whether the respective graphical operation leads to a neighbouring equivalence class are designed in that paper. One of the operations and the respective legality test are aimed to generate upper neighbours, the other operation and test correspond to lower neighbours. Although the graphical description of the inclusion neighbourhood in terms of individual networks from Section 2.4 implies that every inclusion neighbour can be reached in this way the method has two drawbacks.

- The first drawback of the method is that it is tentative: different graphical operations may lead to the same equivalence class. Therefore, additional checking must be done to cure this imperfection.
- The second drawback of this mechanistic approach is that it does not allow one to discern possible internal structure of the inclusion neighbourhood.

Auvray and Wehenkel [2] made an attempt at direct characterization of the inclusion neighbourhood. Their characterization of the upper inclusion neighbourhood, that is, of those neighbouring equivalence classes which describe more independencies, removes the first drawback. They uniquely characterized and classified neighbouring equivalence classes of a given equivalence class (described in terms of the respective essential graph) by means of certain mathematical objects. However, these object are still unnecessarily complicated which means that their characterization of upper inclusion neighbourhood is too awkward. In particular, the second drawback is not removed by their approach since their approach does not allow one to make out the internal structure of the inclusion neighbourhood. Moreover, their characterization is incomplete: only partial direct characterization of lower neighbours is given there.

1.3 Compact Characterization of Inclusion Neighbours

In this paper an elegant characterization of the inclusion neighbourhood of a given equivalence class in terms of the respective essential graph is presented. Each inclusion neighbour is uniquely described by a pair $([a,b],C)$ where $[a,b]$ is an unordered pair of distinct nodes and $C \subseteq N \setminus \{a,b\}$ a disjoint set of nodes. More specifically, $[a,b]$ is an edge of the essential graph in case of the upper neighbourhood while $[a,b]$ is a pair of nodes which is not an edge in the essential graph in case of the lower neigbourhood.

The first new observation made in this paper is that every inclusion neighbour is uniquely characterized by a pair $([a,b],C)$ of this kind. The second observation is that, for given $[a,b]$, the collection of those sets C which correspond to inclusion neighbours has a special form.

In this contribution, a complete analysis of the upper inclusion neighbourhood is given. In this case, the collection of sets C for a given edge $[a,b]$ has the form of a *tuft*. This means that it is a collection of sets with the least set (= the unique minimal set) and with possibly several maximal sets such that every set which contains the least set and which is contained in one of the maximal sets belongs to the collection. In particular, a tuft is completely described by its least set and by the list of its maximal sets. Given an essential graph G^* and an edge $[a,b]$ in G^* the least and maximal sets of the respective tuft of sets are characterized directly in terms of G^*.

The structure of the lower inclusion neighbourhood is similar but more complex. In that case, given a pair of nodes which is not an edge in the essential graph, the respective collection of sets C is the union of (at most) two tufts. The least and maximal sets of these two tufts can also be read from the essential graph. However, because of page limitation the case of lower neighbourhood will be completely analyzed in a future paper [17].

Note that the characterization of inclusion neighbours by means of pairs $([a,b],C)$ where $C \subseteq N \setminus \{a,b\}$ and the way how it is done in this paper is not incidental. An interesting fact is that, from a certain perspective which is explained in details in Chapter 8 of [15], the pair $([a,b],C)$ has close relation to conditional independence interpretation of the 'move' from the considered equivalence class to its respective inclusion neighbour.

The proofs given in the Appendix combine the ideas motivated by an arithmetic approach to the description of Bayesian network models from [15] with certain graphical procedures which were already used in [2].

2 Basic Concepts

2.1 Graphical Notions

Graphs considered here have a finite non-empty set N as the set of *nodes* and two possible types of edges. An undirected edge or a *line* over N is a subset of N of cardinality two, that is, an unordered pair $\{a,b\}$ where $a,b \in N$, $a \neq b$. The respective notation is $a — b$. A directed edge or an *arrow* over N is an ordered pair (a,b) where $a,b \in N$, $a \neq b$. The notation $a \to b$ reflects its pictorial representation. A *hybrid graph* over N is a graph without multiple edges, that is, a triplet $H = (N, \mathcal{L}(H), \mathcal{A}(H))$ where N is a set of nodes, $\mathcal{L}(H)$ a set of lines over N and $\mathcal{A}(H)$ a set of arrows over N such that whenever $(a,b) \in \mathcal{A}(H)$ then $(b,a) \notin \mathcal{A}(H)$ and $\{a,b\} = \{b,a\} \notin \mathcal{L}(H)$. A pair $[a,b]$ of distinct elements of N will be called an *edge in* H (between a and b) if one of the following cases occurs: $a — b$ in H, $a \to b$ in H and $b \to a$ in H. If $\emptyset \neq A \subseteq N$ then the *induced subgraph* H_A of H is the triplet $(A, \mathcal{L}(H) \cap \mathcal{P}(A), \mathcal{A}(H) \cap (A \times A))$ where $\mathcal{P}(A)$ denotes the power set of A (= the collection of subsets of A).

A set $K \subseteq N$ is *complete* in a hybrid graph H over N if $\forall\, a,b \in N$ $a \neq b$ one has $a — b$ in H. By a *clique* of H will be understood a maximal complete set in H (with respect to set inclusion). The collection of cliques of H will be denoted by *cliques*(H). A set $C \subseteq N$ is *connected* in H if, for every $a,b \in N$, there exists an undirected path connecting them, that is, a sequence of distinct nodes $a = c_1, \ldots, c_n = b$, $n \geq 1$ such that $c_i — c_{i+1}$ in H for $i = 1, \ldots, n-1$. Connectivity *components* of H are maximal connected sets in H.

An *undirected graph* is a hybrid graph without arrows, that is, $\mathcal{A}(H) = \emptyset$. An undirected graph H is *triangulated* if, for every undirected cycle in H which has the length at least four, that is, for any sequence $c_1, \ldots, c_n, c_{n+1} = c_1$, $n \geq 4$ where $c_1, \ldots, c_n$ are distinct and $c_i — c_{i+1}$ in H for $i = 1, \ldots, n$, there exists a *chord* in H, that is, an edge $c_i — c_j$ where $1 \leq i, j \leq n$ and $1 < j - i < n - 1$.

A *directed graph* is a hybrid graph having arrows only, that is, $\mathcal{L}(H) = \emptyset$. An *acyclic directed graph* is a directed graph without directed cycles, that is, without any sequence $d_1, \ldots, d_n, d_{n+1} = d_1$, $n \geq 3$ such that $d_1, \ldots, d_n$ are distinct and $d_i \to d_{i+1}$ in H for $i = 1, \ldots, n$. The well-known fact is that a directed graph is acyclic iff there exists a total ordering of all nodes of N $a_1, \ldots, a_m$, $m \geq 1$ which is *consistent* with the direction of arrows, that is, whenever $a_i \to a_j$ in H then $i < j$.

A *chain graph* is a hybrid graph H for which there exists a *chain*, that is, an ordered partitioning of N into non-empty sets, called *blocks*, $B_1, \ldots, B_m$, $m \geq 1$ such that

- if $a — b$ in H then $a, b \in B_i$ for some $1 \leq i \leq m$,
- if $a \to b$ in H then $a \in B_i, b \in B_j$ with $1 \leq i < j \leq m$.

Clearly, every undirected graph and every acyclic directed graph is a chain graph. An equivalent definition of a chain graph is that it is a hybrid graph H without *semi-directed cycles*, that is, without any sequence $d_1, \ldots, d_n, d_{n+1} = d_1$, $n \geq 3$ such that $d_1, \ldots, d_n$ are distinct, $d_1 \to d_2$ in H and $\forall\, i = 2, \ldots, n$ either $d_i \to d_{i+1}$ or $d_i — d_{i+1}$ in H - see Lemma 2.1 in [13]. In particular, there is no arrow in a chain graph between nodes of a connected set $C \subseteq N$; in other words, the induced subgraph H_C is undirected. Thus, the set of *parents* of C, that is,

$$pa_H(C) = \{a \in N\,;\, \exists\, b \in C \quad a \to b \text{ in } H\,\}$$

is disjoint with C. The set

$$ne_H(C) = \{a \in N \setminus C\,;\, \exists\, b \in C \quad a — b \text{ in } H\,\}$$

will be named the set of *neighbours* of C.

2.2 Bayesian Networks and Their Equivalence

A *Bayesian network* is a certain statistical model, that is, a class of (multidimensional probability) distributions, appended to an acyclic directed graph. It could be introduced as the class of distributions (on a fixed sample space) which factorize according to the graph in a certain way. An alternative definition of that class can be given terms of conditional independence restrictions, using the *d*-separation criterion from [11] or using the moralization criterion from [8] which are known to be equivalent. Since exact definitions of these concepts are not needed in this paper they are omitted. Nevertheless, given an acyclic directed graph G over N, the symbol $\mathcal{I}(G)$ will be used to denote the collection of conditional independence restrictions determined by G. Moreover, the phrase "Bayesian network" will be used as a synonym for an acyclic directed graph throughout the rest of the paper.

An important concept is the concept of *equivalence of Bayesian networks.* Two Bayesian networks G_1 and G_2 are considered to be equivalent if they represent the same statistical model, which requirement is typically equivalent to the condition $\mathcal{I}(G_1) = \mathcal{I}(G_2)$. Given a equivalence class $\mathcal{G}$ of Bayesian networks over N the symbol $\mathcal{I}(\mathcal{G})$ will denote the shared collection of conditional independence restrictions $\mathcal{I}(G)$ for $G \in \mathcal{G}$. Verma and Pearl [18] gave a direct graphical characterization of equivalent Bayesian networks which can be used as its formal definition here. The *underlying graph* of a hybrid graph H over N is an undirected graph H^u over N such that a — b in H^u iff $[a, b]$ is an edge in H. An *immorality* in H is an induced subgraph of H shown in the left-hand picture of Figure 1, that is, the configuration $a \to c \leftarrow b$ where a, b, c are distinct nodes and the pair $[a, b]$ is not an edge in H. Two Bayesian networks G_1, G_2 over N are (graphically) *equivalent* iff they have the same underlying graph and the same collection of immoralities. The equivalence characterization makes the following definition consistent: given an equivalence class $\mathcal{G}$ of Bayesian networks, a pair $[a, b]$ of distinct nodes is called an *edge in* $\mathcal{G}$ if $[a, b]$ is an edge in some $G \in \mathcal{G}$, which means, it is an edge in every $G \in \mathcal{G}$.

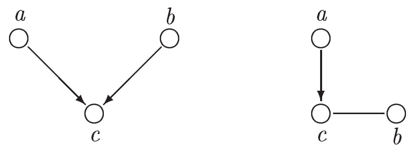

Fig. 1. An immorality and a flag.

2.3 Essential Graphs

An equivalence class $\mathcal{G}$ of Bayesian networks (over N) can be described by its *essential graph* which is a hybrid graph G^* (over N) such that

- $a \to b$ in G^* if and only if $a \to b$ in G for every $G \in \mathcal{G}$,
- a — b in G^* if and only if there exist $G_1, G_2 \in \mathcal{G}$ such that $a \to b$ in G_1 and $b \to a$ in G_2.

Example 1. Let us consider the equivalence class $\mathcal{G}$ of Bayesian networks over $N = \{a, b, c, d\}$ shown on the left-hand side of Figure 2. The respective essential graph G^* is in the right-hand picture of the same figure.

A graphical characterization of essential graphs was given by Andersson, Madigan and Perlman as Theorem 4.1 in [1]. Recently, a simpler alternative characterization has been found in [16] and, independently, in [12]. As complete characterization of essential graphs is not needed in this paper, it is omitted.

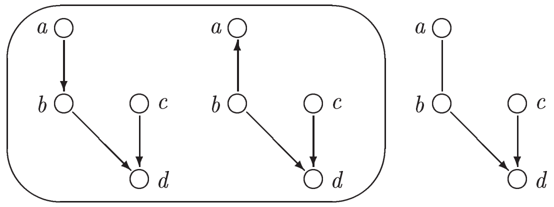

Fig. 2. An equivalence class of Bayesian networks and the respective essential graph.

However, what is needed is the following observation. It follows from Theorem 4.1 of [1] that every essential graph H (of an equivalence class of Bayesian networks) is a chain graph without flags such that, for every component C of H, the induced subgraph H_C is a triangulated graph. Recall that by a *flag* in a hybrid graph H is meant an induced subgraph of H shown in the right-hand picture of Figure 1, that is, the configuration $a \rightarrow c - b$ where a, b, c are distinct nodes and the pair $[a, b]$ is not an edge in H. Note that every chain graph without flags has the following property: for every component C of H and $a, b \in C$ one has $pa_H(a) = pa_H(b)$; in particular, $pa_H(a) = pa_H(C)$ for any $a \in C$.

To formulate another important fact about the essential graph recall that every chain graph can also be interpreted as a statistical model. Indeed, the moralization criterion from [8], alternatively the c-separation criterion from [14], allows one to ascribe the collection of conditional independence restrictions $\mathcal{I}(H)$ to every chain graph H over N. This defines the respective statistical model and induces the concept of equivalence for chain graphs over N. A direct graphical characterization of equivalent chain graphs was given by Frydenberg [6]. It follows from that result that two chain graphs without flags are equivalent iff they have the same underlying graph and immoralities - see Lemma 2 in [16]. Thus, Frydenberg's result implies the result by Verma and Pearl [18]. Another basic fact about the essential graph G^* of an equivalence class $\mathcal{G}$ is that it is equivalent to every $G \in \mathcal{G}$, that is, $\mathcal{I}(G^*) = \mathcal{I}(G)$ for $G \in \mathcal{G}$ - see Corollary 2 in [16].

Remark 1. Note that there are other possible ways of representing an equivalence class $\mathcal{G}$ of Bayesian networks. One of them is the concept of *largest chain graph* [6] of the collection of chain graphs which are equivalent to (any) $G \in \mathcal{G}$. Another alternative is brought by the arithmetic approach presented in Section 8.4 of [15] which offers the concept of *standard imset* - see Section 5.1 in the Appendix.

2.4 Inclusion Ordering and Neighbourhood

The *inclusion ordering* on the set of equivalence classes of Bayesian networks (over N) is defined by the binary relation $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}(\mathcal{L})$ for equivalence classes $\mathcal{K}$ and $\mathcal{L}$. The symbol $\mathcal{I}(\mathcal{K}) \subset \mathcal{I}(\mathcal{L})$ will denote the *strict ordering*, that is, the situation when $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}(\mathcal{L})$ and $\mathcal{I}(\mathcal{K}) \neq \mathcal{I}(\mathcal{L})$. Finally, the symbol $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$ will mean that $\mathcal{I}(\mathcal{K}) \subset \mathcal{I}(\mathcal{L})$ but there is no equivalence class $\mathcal{G}$ of Bayesian networks (over N) such that $\mathcal{I}(\mathcal{K}) \subset \mathcal{I}(\mathcal{G}) \subset \mathcal{I}(\mathcal{L})$. If this is the case then $\mathcal{L}$ is called the *upper neighbour* of $\mathcal{K}$ and $\mathcal{K}$ is called the *lower neighbour* of $\mathcal{L}$. By the *inclusion neighbourhood* of an equivalence class is understood the collection its upper and lower neigbours.

Transformational characterization of the inclusion ordering from [5] allows one to derive a simple graphical description of the relation $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$ as a consequence - see Lemma 8.5 in [15].

Lemma 1. *If $\mathcal{K}$ and $\mathcal{L}$ are equivalence classes of Bayesian networks over N then one has $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$ iff there exists $K \in \mathcal{K}$ and $L \in \mathcal{L}$ such that L is made of K by the removal of (exactly) one edge.*

Remark 2. The relation $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}(\mathcal{L})$ corresponds to the situation when the statistical model given by $K \in \mathcal{K}$ contains the statistical model given by $L \in \mathcal{L}$. The networks in $\mathcal{K}$ have more edges than networks in $\mathcal{L}$ then. The reader may ask why $\mathcal{L}$ is supposed to be 'above' $\mathcal{K}$ in this paper (and not conversely). The terminology used in this paper simply emphasizes the conditional independence interpretation of considered statistical models which is in the center of author's interests - for more detailed justification see Remark 8.10 in [15].

2.5 Tuft of Sets

Let $\mathcal{T}$ be a non-empty collection of subsets of N, that is, $\emptyset \neq \mathcal{T} \subseteq \mathcal{P}(N)$, and $\mathcal{T}_{\max}$ denotes the collection of maximal sets in $\mathcal{T}$ (with respect to set inclusion). The collection $\mathcal{T}$ will be called a *tuft* of sets if

- $\mathcal{T}$ has the least set $T_{\min}$, that is, $T_{\min} \in \mathcal{T}$ with $T_{\min} \subseteq T$ for $T \in \mathcal{T}$,
- every set $T \subseteq N$ such that $T_{\min} \subseteq T \subseteq T'$ for some $T' \in \mathcal{T}_{\max}$ belongs to $\mathcal{T}$.

Thus, a tuft of sets $\mathcal{T}$ is determined by its unique least set $T_{\min}$ and by the class of its maximal sets $\mathcal{T}_{\max}$. Alternatively, it can be described by $T_{\min}$ and the class $\{T' \setminus T_{\min}\,;\, T' \in \mathcal{T}_{\max}\}$. More specifically, assume that $A \subseteq N$ and $\mathcal{B}$ is a non-empty class of incomparable subsets of $N \setminus A$, that is, there are no $B, B' \in \mathcal{B}$ with $B \subset B'$. Introduce a special notation:

$$\mathsf{TUFT}(A|\mathcal{B}) \,=\, \{\, T = A \cup C\,;\, \exists\, B \in \mathcal{B} \;\; C \subseteq B\,\}\,.$$

Evidently, $\mathcal{T} = \mathsf{TUFT}(A|\mathcal{B})$ is a tuft such that $\mathcal{T}_{\max} = \{A \cup B; B \in \mathcal{B}\}$ and $T_{\min} = A$. Of course, every tuft of subsets of N can be described in this way.

Example 2. Suppose $N = \{a, b, c, d\}$ and put $A = \{a\}$. Consider the class $\mathcal{B} = \{\,\{b\}, \{c\}, \{d\}\,\}$ which is a class of incomparable subsets of $N \setminus A$ (actually, the sets in $\mathcal{B}$ are disjoint). Then $\mathsf{TUFT}(A|\mathcal{B})$ consists of four sets: $\{a\}$, $\{a, b\}$, $\{a, c\}$ and $\{a, d\}$. The tuft is shown in Figure 3.

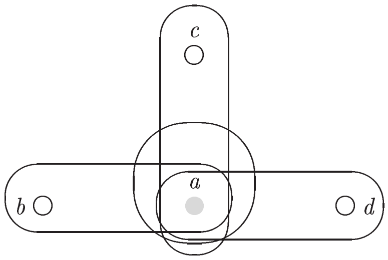

Fig. 3. A tuft of sets $\mathsf{TUFT}(\{a\}|\{b\},\{c\},\{d\})$.

3 Upper Inclusion Neigbourhood

3.1 Description of Upper Neighbours

By the *upper neigbouhood* of an equivalence class $\mathcal{K}$ of Bayesian networks is understood the collection $o^{\uparrow}(\mathcal{K})$ of equivalence classes $\mathcal{L}$ such that $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$. It follows from Lemma 1 that each $K \in \mathcal{K}$ and each edge in $\mathcal{K}$ define together an element of $o^{\uparrow}(\mathcal{K})$ and every element of $o^{\uparrow}(\mathcal{K})$ is obtained in this way. Thus, the upper neigbourhood $o^{\uparrow}(\mathcal{K})$ is, in fact, described in terms of elements of $\mathcal{K}$. Nevertheless, the above described correspondence is not a one-to-one mapping because different elements of $\mathcal{K}$ may yield the same neighbouring class $\mathcal{L}$.

One the other hand, every neighbouring class is uniquely characterized by a certain pair $([a,b],C)$ where $a,b \in N$, $a \neq b$ and $C \subseteq N \setminus \{a,b\}$. The pair $([a,b],C)$ can be introduced in graphical terms as follows.

> Let $\mathcal{K}$ be an equivalence class of Bayesian networks over N, $\mathcal{L} \in o^{\uparrow}(\mathcal{K})$. Choose $K \in \mathcal{K}$ and $L \in \mathcal{L}$ such that L is obtained from K by the removal of an arrow $a \to b$ in K. Then $\mathcal{L}$ will be described by the pair $([a,b],C)$ where $C = pa_K(b) \setminus \{a\}$.

To show that the definition above is consistent one has to show that the pair $([a,b],C)$ does not depend on the choice of K and L and that distinct pairs are ascribed to distinct upper neighbours.

Proposition 1. *Let $\mathcal{K}$ be an equivalence class of Bayesian networks, $\mathcal{L}_1,\mathcal{L}_2 \in o^{\uparrow}(\mathcal{K})$. Suppose, for $i=1,2$, that graphs $K_i \in \mathcal{K}$ and $L_i \in \mathcal{L}_i$ are given such that L_i is made of K_i by the removal of an arrow $a_i \to b_i$ in K_i and $C_i = pa_{K_i}(b_i) \setminus \{a_i\}$. Then $\mathcal{L}_1 = \mathcal{L}_2$ iff $[\{a_1,b_1\} = \{a_2,b_2\}$ and $C_1 = C_2]$.*

The proof of Proposition 1, which is given in Section 5.1 of the Appendix, is based on a special arithmetic characterization of equivalence of Bayesian networks. Note that one can perhaps also prove this result using purely graphical tools, but the given proof is more elegant.

Example 3. To illustrate concepts introduced above let us consider an equivalence class $\mathcal{K}$ of Bayesian networks over $N = \{a, b, c, d, e\}$ shown in the lower layer of Figure 4. For every $i = 1, 2, 3$, an acyclic directed graph L_i is obtained from $K_i \in \mathcal{K}$ by the removal of the arrow $a \to b$ (see the medium layer of the figure). In this example, each of K_i, $i = 1, 2, 3$ establishes a different neighbouring class $\mathcal{L}_i \in o^{\uparrow}(\mathcal{K})$ - the respective essential graphs are in the upper layer of Figure 4. Since $pa_{K_2}(b) \setminus \{a\} = \{c, d\}$, the equivalence class containing L_2 is characterized by the pair $([a, b], \{c, d\})$.

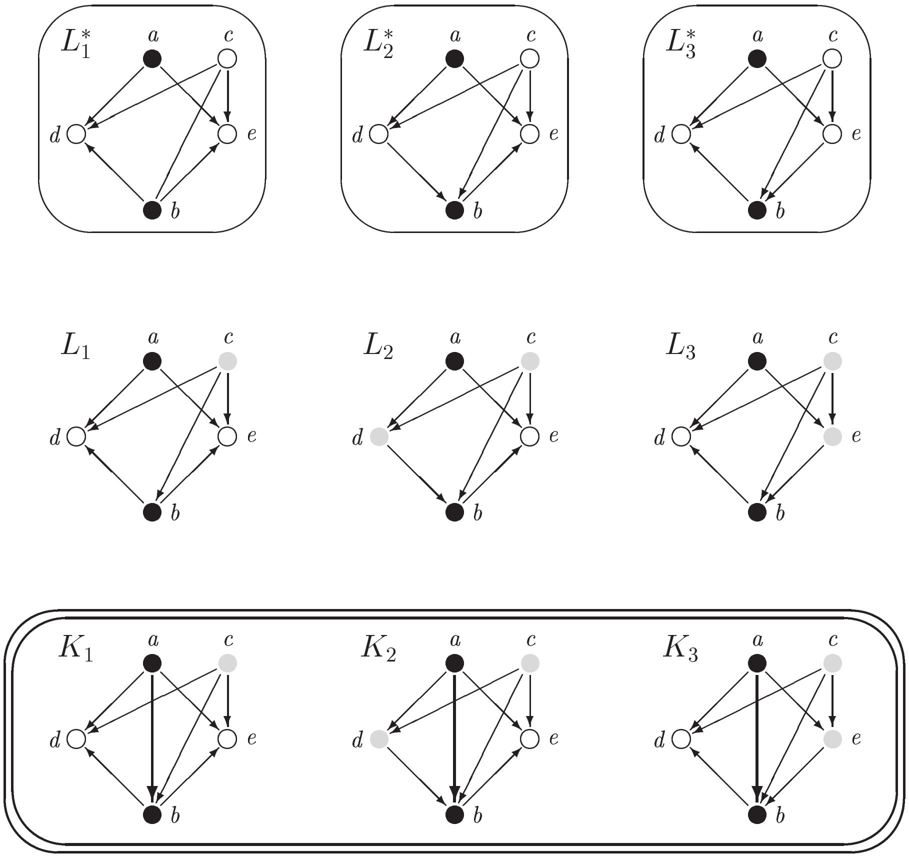

Fig. 4. Equivalence class of Bayesian networks and some of its upper neighbours.

Remark 3. The pair $([a, b], C)$ describing uniquely an upper neighbour $\mathcal{L} \in o^{\uparrow}(\mathcal{K})$ was introduced in terms of individual networks from $\mathcal{K}$ and $\mathcal{L}$. If $\mathcal{K}$ and $\mathcal{L}$ are represented by the respective essential graphs K^* and L^* then $[a, b]$ is simply the edge of K^* which is not an edge in L^*. The question of how to define C in terms of K^* and L^* have not been examined so far by the author. On the other hand, the pair $([a,b],C)$ is obtained immediately if $\mathcal{K}$ and $\mathcal{L}$ are represented by means of their standard imsets - see Section 5.1 in the Appendix.

3.2 Characterization of Upper Neighbourhood

Given an equivalence class $\mathcal{K}$ the next step is to characterize those pairs $([a,b],C)$ which define elements $L \in o^{\uparrow}(\mathcal{K})$. In this section, this task is answered for a fixed unordered pair of distinct nodes $[a,b]$. For this purpose, put

$$\mathcal{C}_{\mathcal{K}}^{-}(a \to b) \;=\; \{\, C\,;\; \exists\, K \in \mathcal{K} \text{ such that } a \to b \text{ in } K \text{ and } C = pa_K(b) \setminus \{a\}\,\}$$

for every ordered pair of distinct nodes (a,b). It follows from what it says in Section 3.1 that $\mathcal{C}_{\mathcal{K}}^{-}(a \to b) \cup \mathcal{C}_{\mathcal{K}}^{-}(b \to a)$ is the class of sets which has to be characterized. Therefore, given (a,b), one needs to find out when $\mathcal{C}_{\mathcal{K}}^{-}(a \to b)$ is non-empty and describe that collection if it is non-empty.

Proposition 2. *Let $\mathcal{K}$ be an equivalence class of Bayesian networks, K^* the essential graph of $\mathcal{K}$ and (a,b) an ordered pair of distinct nodes of K^*. Put $P = pa_{K^*}(b) \setminus \{a\}$ and $M = \{c \in ne_{K^*}(b) \setminus \{a\}\,;\;\; [a,c] \text{ is an edge in } K^*\}$. Then*

(i) *$\mathcal{C}_{\mathcal{K}}^{-}(a \to b) \neq \emptyset$ iff $[a,b]$ is an edge in K^* and $b \notin pa_{K^*}(a)$, that is, either $a \to b$ in K^* or $a — b$ in K^*.*
(ii) *If $a \to b$ in K^* or $a — b$ in K^* then $\mathcal{C}_{\mathcal{K}}^{-}(a \to b) = \mathsf{TUFT}(P|cliques(K^*_M))$ where $cliques(K^*_{\emptyset}) = \{\emptyset\}$ by convention.*

The proof is given in Section 5.2 of the Appendix.

Corollary 1. *Let $\mathcal{K}$ be an equivalence class of Bayesian networks over N, K^* the essential graph of $\mathcal{K}$ and $[a,b]$ an edge in K^*. Then the collection of those sets $C \subseteq N \setminus \{a,b\}$ such that $([a,b],C)$ describes an upper neighbour $\mathcal{L} \in o^{\uparrow}(\mathcal{K})$ is a tuft $\mathsf{TUFT}(P|cliques(K^*_M))$ where*

(a) *if $a \to b$ in K^* then $P = pa_{K^*}(b) \setminus \{a\}$ and $M = ne_{K^*}(b)$,*
(b) *if $a \leftarrow b$ in K^* then $P = pa_{K^*}(a) \setminus \{b\}$ and $M = ne_{K^*}(a)$,*
(c) *if $a — b$ in K^* then $M = \{c \in N; a — c — b \text{ in } K^*\}$ and $P = pa_{K^*}(b) = pa_{K^*}(a)$.*

Proof. Recall that one needs to characterize $\mathcal{C}_{\mathcal{K}}^{-}(a \to b) \cup \mathcal{C}_{\mathcal{K}}^{-}(b \to a)$ using Proposition 2. In case (a) observe that $\mathcal{C}_{\mathcal{K}}^{-}(b \to a) = \emptyset$. As concerns $\mathcal{C}_{\mathcal{K}}^{-}(a \to b)$, the fact that K^* has no flags implies $M = ne_{K^*}(b)$. The case (b) is symmetric. In case (c) realize that the fact that K^* has no flags implies $pa_{K^*}(b) = pa_{K^*}(a)$. This observation and the fact that K^* is a chain graph allow one to show that $\mathcal{C}_{\mathcal{K}}^{-}(a \to b) = \mathcal{C}_{\mathcal{K}}^{-}(b \to a) \neq \emptyset$.

Example 4. To illustrate the previous result consider the essential graphs shown in Figure 5. The case (a) from Corollary 1 occurs for the graph K^* in the left-hand picture of the figure. More specifically, one has $P = \{c\}$, $M = \{d, e\}$ and $cliques(K^*_M) = \{\,\{d\}, \{e\}\,\}$. Thus, the class of sets $C \subseteq N \setminus \{a, b\}$ such that $([a, b], C)$ describes an upper neighbour of the respective equivalence class is $\mathsf{TUFT}(\{c\}|\{d\}, \{e\})$, that is, the class which involves three sets: $\{c\}$, $\{c, d\}$ and $\{c, e\}$. Indeed, it was shown in Example 3 that those upper neighbours of the respective equivalence class which "correspond" to the "removal" of $a \to b$ are characterized by pairs $([a, b], \{c\})$, $([a, b], \{c, d\})$ and $([a, b], \{c, e\})$.

If the graph G^* in the right-hand picture of Figure 5 is considered then the case (c) from Corollary 1 occurs. One has $P = \emptyset$, $M = \{c, d, e\}$ and $cliques(G^*_M) = \{\,\{c, d\}, \{d, e\}\,\}$. The respective class $\mathsf{TUFT}(\emptyset|\{c, d\}, \{d, e\})$ has five sets: $\emptyset$, $\{c\}$, $\{d\}$, $\{e\}$, $\{c, d\}$ and $\{d, e\}$.

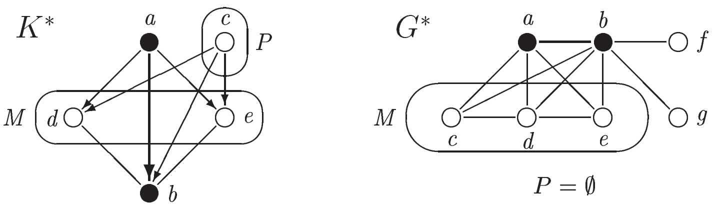

Fig. 5. Two essential graphs.

4 Conclusions

In this contribution a characterization of the upper inclusion neighbourhood was presented. Analogous results for the lower inclusion neighbourhood have also been achieved and they will be presented in a later paper [17]. Note that lower neighbours of a given equivalence class of Bayesian networks over N can also be described uniquely by pairs $([a, b], C)$ where $a, b \in N$, $a \neq b$ and $C \subseteq N \setminus \{a, b\}$ and that, given $[a, b]$, the class of sets C which correspond to lower neighbours is the union of two tufts which can be characterized in terms of the respective essential graph. An important fact is that if $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$ then the pair $([a, b], C)$ describing $\mathcal{L}$ as one of the upper neighbours of $\mathcal{K}$ coincides with the pair $([a, b], C)$ which describes $\mathcal{K}$ as one of the lower neighbours of $\mathcal{L}$. Thus, there is internal consistency of both characterizations and the pair $([a, b], C)$ can be viewed as a natural characteristic of the 'move' between $\mathcal{K}$ and $\mathcal{L}$.

As explained in Section 1 the presented characterization is more elegant than the previous ones. Indeed, Chickering [5] only gave a tentative algorithmic method and Auvrey and Wehenkel [2] characterized every inclusion neighbour by an unordered pair of nodes and by an opaque collection of immoralities, namely those which are either created or cancelled if an equivalence class is replaced by its inclusion neighbour. Finally, the characterization presented in this paper has a close connection to an arithmetic method of description of equivalence clases of Bayesian networks (from Chapter 8 of [15]) and leads to conditional independence interpretation of 'moves' in the method of local search.

References

1. Andersson, S. A., Madigan, D. and Perlman, M. D.: A characterization of Markov equivalence classes for acyclic digraphs. *Annals of Statistics* 25 (1997) 505–541.
2. Auvray, V. and Wehenkel, L.: On the construction of the inclusion boundary neighbourhood for Markov equivalence clases of Bayesian network structures. In: Darwiche, A. and Friedman, N. (eds.): Uncertainty in Artificial Intelligence 18. Morgan Kaufmann, San Francisco (2002) 26–35.
3. Bouckaert, R. R.: Bayesian belief networks: from construction to inference. PhD thesis, University of Utrecht (1995).
4. Castelo, R.: The discrete acyclic digraph Markov model in data mining. PhD thesis, University of Utrecht (2002).
5. Chickering, D. M.: Optimal structure identification with greedy search. To appear in *Journal of Machine Learning Research* (2003).
6. Frydenberg, M.: The chain graph Markov property. *Scandinavian Journal of Statistics* 17 (1990) 333–353.
7. Kočka, T. and Castello, R.: Improved learning of Bayesian networks. In: Breese, J. and Koller, D. (eds.): Uncertainty in Artificial Intelligence 17. Morgan Kaufmann, San Francisco (2001) 269–276.
8. Lauritzen, S. L.: Graphical Models. Clarendon Press, Oxford (1996).
9. Meek, C.: Causal inference and causal explanation with background knowledge. In: Besnard, P. and Hanks, S. (eds.): Uncertainty in Artificial Intelligence 11. Morgan Kaufmann, San Francisco (1995) 403–410.
10. Meek, C.: Graphical models, selectin causal and statistical models. PhD thesis, Carnegie Mellon University (1997).
11. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo (1988).
12. Rovenato, A.: A unified approach to the characterisation of equivalence classes of DAGs, chain graphs with no flags and chain graphs. Technical report, Dipartimento di Scienze Sociali Cognitive e Quantitative, University of Modena and Reggio Emilia (2003), submitted to *Annals of Statistics*.
13. Studený, M.: A recovery algorithm for chain graphs. *International Journal of Approximate Reasoning* 17 (1997) 265–293.
14. Studený, M. and Bouckaert, R. R.: On chain graph models for description of conditional independence structures. *Annals of Statistics* 26 (1998) 1434–1495.
15. Studený, M.: On probabilistic conditional independence structures. A research monograph, Institute of Information Theory and Automation, Prague, Czech Republic (2003), offerred to *Springer Verlag*.
16. Studený, M.: Characterization of essential graphs by means of an operation of legal merging of components. Submitted to *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*.
17. Studený, M.: Characterization of inclusion neighbourhood in terms of the essential graph: lower neighbours. In preparation, to be submitted to Proceedings of WUPES 2003.

18. Verma, T. and Pearl, J.: Equivalence and synthesis of causal models. In: Bonissone, P. P., Henrion, M., Kanal L. N. and Lemmer, J. F. (eds.): Uncertainty in Artificial Intelligence 6. Elsevier, New York (1991) 220–227.

5 Appendix

5.1 Algebraic Approach

Special concepts from [15] are recalled in this section. By an *imset* over N will be understood any integer-valued function on $\mathcal{P}(N)$. Algebraic operations with imsets are defined coordinatewisely. Given $A \subseteq N$ the symbol δ_A will denote a special imset which identifies the set A:

$$\delta_A(B) = \begin{cases} 1 & \text{if } B = A, \\ 0 & \text{if } B \neq A, \end{cases} \qquad \text{for any } B \subseteq N\,.$$

If G is a Bayesian network over N then the *standard imset* for G is given by the formula

$$u_G \;=\; \delta_N - \delta_\emptyset + \sum_{i \in N} \{\, \delta_{pa_G(i)} - \delta_{\{i\} \cup pa_G(i)} \,\}\,. \tag{1}$$

The following result is proved as Corollary 7.1 in [15].

Theorem 1. *Bayesian networks K and L over N are equivalent iff $u_K = u_L$.*

In particular, a standard imset $u_{\mathcal{G}}$ for any equivalence class $\mathcal{G}$ of Bayesian networks can be introduced as the shared standard imset u_G for $G \in \mathcal{G}$. Note that there exists a formula for $u_{\mathcal{G}}$ in terms of the essential graph. A basic observation is as follows.

Lemma 2. *Suppose that K, L are Bayesian networks over N and L is made of K by the removal of an arrow $a \to b$ in K. Then $u_L - u_K = \delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C} + \delta_{\{a,b\} \cup C}$ where $C = pa_K(b) \setminus \{a\}$.*

Proof. It follows from the assumption that $pa_K(c) = pa_L(c)$ for every $c \in N \setminus \{b\}$. Thus, after the substitution of (1) into $u_L - u_K$ most of the terms are cancelled. As $pa_L(b) = C$ and $pa_K(b) = \{a\} \cup C$ the result is what Lemma 2 says.

Remark 4. Note that the inclusion neighbourhood relation can also be characterized in terms of standard imsets. It is shown in [15] (Corollary 8.4) that, for equivalence classes $\mathcal{K}$ and $\mathcal{L}$, one has $\mathcal{I}(\mathcal{K}) \sqsubset \mathcal{I}(\mathcal{L})$ iff $u_{\mathcal{L}} - u_{\mathcal{K}}$ is an imset of the form $\delta_C - \delta_{\{a\} \cup C} - \delta_{\{b\} \cup C} + \delta_{\{a,b\} \cup C}$ for some $a, b \in N$, $a \neq b$ and $C \subseteq N \setminus \{a, b\}$. The inclusion ordering can be characterized analogously (see Lemma 8.6 in [15]).

The proof of Proposition 1 can be done on basis of Lemma 2 easily.

Proof. If the assumption of Proposition 1 are fulfilled then $u_{K_1} = u_{K_2}$ by Theorem 1. Moreover, $\mathcal{L}_1 = \mathcal{L}_2$ iff $u_{L_1} = u_{L_2}$ which is equivalent to $u_{L_1} - u_{K_1} = u_{L_2} - u_{K_2}$. However, the latter is equivalent to

$$\delta_{C_1} - \delta_{\{a_1\} \cup C_1} - \delta_{\{b_1\} \cup C_1} + \delta_{\{a_1, b_1\} \cup C_1} \;=\; \delta_{C_2} - \delta_{\{a_2\} \cup C_2} - \delta_{\{b_2\} \cup C_2} + \delta_{\{a_2, b_2\} \cup C_2}$$

by Lemma 2. This is nothing but the condition that $C_1 = C_2$ and $\{a_1, b_1\} = \{a_2, b_2\}$.

5.2 Bayesian Network Construction

Given an undirected graph H over N with $|N| = n$, by *perfect numbering* is meant a total ordering $a_1, \ldots, a_n$ of its nodes such that $ne_H(a_i) \cap \{a_j; j < i\}$ is a complete set in H for every $i = 1, \ldots, n$. It is a well-known fact that every triangulated graph admits a perfect numbering. More specifically, the following statement is true.

Lemma 3. *Let H be a triangulated graph, $A \subseteq N$ a complete set in H and $a_1, \ldots, a_r$, $r \geq 0$ any ordering of nodes in A. Then there exists a perfect numbering of nodes of H which starts by the sequence $a_1, \ldots, a_r$.*

Proof. The required perfect numbering is constructed in the reverse order by means of repeated application of well-known Dirac's lemma - see Lemma 2.9 in [8]. Recall that a node is *simplicial* in an undirected graph H' if $ne_{H'}(a)$ is complete in H'. Dirac's lemma says that if H' is a triangulated graph over N' and N' is not complete in H' then H' has two simplicial nodes $a, b \in N'$ such that $[a, b]$ is not an edge in H'.

Put $H^n = H'$. If N is complete in H^n then the claim of Lemma 3 is evident. If this is not the case then the assumption that A is complete in H^n implies by Dirac's lemma that a simplicial node $a_n \in N \setminus A$ exists. Put $H^{n-1} = H^n_{N\setminus\{a_n\}}$ and observe that H^{n-1} is a triangulated graph and A is complete in H^{n-1}. The procedure can be repeated untill the set of nodes of some H^j for $1 \leq j \leq n$ is complete in H^j.

The preceding statement makes it possible to construct a Bayesian network on basis of an essential graph.

Lemma 4. *Let H be a chain graph without flags such that, for every component C of H, the graph H_C is triangulated. Moreover, suppose that a perfect numbering of nodes of H_C is prescribed for every component C. Let G be a directed graph made of H in such a way that every line $a — b$ in H is replaced by an arrow $a \to b$ in G provided that a precedes b in the prescribed perfect numbering of the component containing $\{a, b\}$. Then G is a Bayesian network which is equivalent to H.*

Proof. To show that G is acyclic a total ordering of nodes consistent with the direction of arrows is constructed. First, since H is a chain graph one can find a chain $C_1, \ldots, C_m$, $m \geq 1$ for H whose blocks are components of H. Second, within each component, nodes are ordered according to the prescribed numbering. Thus, G is acyclic and both G and H are chain graphs without flags with the same underlying graph.

To evidence that they are equivalent one needs to verify that they have the same immoralities (see Section 2.3). By construction, every immorality in H remains in G. Thus, consider an immorality $a \to c \leftarrow b$ in G and show that it is an immorality in H. First, the cases $a \leftarrow c$ in H and $c \to b$ in H are excluded due to the definition of G. Second, the alternative $a — c — b$ in H also cannot occur because the prescribed numberings are perfect: otherwise a, b, c belong to the same component of H and $a, b \in pa_G(c)$ means that $a, b \in ne_H(c)$ and a, b precede c in the prescribed numbering which implies a contradictory conclusion that $[a, b]$ is an edge in H. Third, the alternatives $a \to c — b$ in H and $a — c \leftarrow b$ in H are impossible because H has no flags. This necessitates $a \to c \leftarrow b$ in H which was desired.

Corollary 2. *Let H be an essential graph, (a, b) an ordered pair of nodes in H such that either $a \to b$ in H or $a — b$ in H. Put $P = \mathrm{pa}_H(b) \setminus \{a\}$ and introduce the set $M = \{c \in \mathrm{ne}_H(b) \setminus \{a\}\,;\ [a, c]$ is an edge in $H\}$. Suppose that $X \subseteq M$ is complete in H. Then there exists a Bayesian network G which is equivalent to H such that $a \to b$ in G and $\mathrm{pa}_G(b) \setminus \{a\} = P \cup X$.*

Proof. Every essential graph is a chain graph which satisfies the assumptions of Lemma 4 (see Section 2.3). Let C denote the component of H which contains b. Evidently $X \cup \{b\} \subseteq M \cup \{b\} \subseteq C$ and $X \cup \{b\}$ is a complete set in H_C. Moreover, if $a — b$ in H then $a \in C$ and $X \cup \{a, b\}$ is a complete set in H_C. By Lemma 3 there exists a perfect numbering of nodes of H_C in which the nodes of X, respectively the nodes of $X \cup \{a\}$, precede the node b and b precedes the remaining nodes of C. After that, arbitrary perfect numberings of nodes of remaining components of H are chosen (by Lemma 3) and Lemma 4 is applied to get the respective Bayesian network G. It follows from the construction that $pa_G(b) = \{a\} \cup P \cup X$.

Now, the proof of Proposition 2 follows.

Proof. The necessity of the condition in (i) follows from the definition of essential graph. Its sufficiency follows from Corollary 2 which says (see Sections 2.3 and 2.5) $P \in \mathsf{TUFT}(P|\mathit{cliques}(K^*_M)) \subseteq \mathcal{C}^-_{\mathcal{K}}(a \to b)$.

The last step is to verify $\mathcal{C}^-_{\mathcal{K}}(a \to b) \subseteq \mathsf{TUFT}(P|\mathit{cliques}(K^*_M))$. If $K \in \mathcal{K}$ such that $a \to b$ in K and $C = pa_K(b) \setminus \{a\}$ then $P \subseteq C$ by the definition of essential graph. Put $X = C \setminus P$ and observe that $X \subseteq M$. Indeed, for any $c \in X \subseteq C \setminus P$ one has $a \to b \leftarrow c$ in K. If $[a, c]$ is not an edge in K then $a \to b \leftarrow c$ is an immorality in K and, therefore, in K^* which contradicts the assumption $c \notin P$. Thus, $[a, c]$ is an edge in K^* which implies $c \in M$ (see the definition of M in Proposition 2). Finally, observe that X is complete as otherwise distinct $c, d \in X$ exist such that $c \to b \leftarrow d$ is an immorality in K and, therefore, in K^* which implies a contradictory conclusion $c, d \in P$.