Prague Stochastics'98

COMPLEXITY OF STRUCTURAL MODELS

Milan Studený[1]

Institute of Information Theory and Automation
Academy of Sciences of Czech Republic
Pod vodárenskou věží 4, 182 08 Prague, Czech Republic

Abstract

Complexity of a model of conditional independence structure is introduced as the least cardinality of a generating set. Four basic types of complexity are distinguished which depend on the type of generating. A method of calculation of complexity of a given conditional independence model is proposed. The method is based on a more effective way of representation of the model by means of a list of dominant conditional independence statements. Prospects of the proposed approach are discussed in the end.

AMS classification: 68T30, 62H05.
Keywords: probabilistic model, semi-graphoid, complexity, dominant triplet.

1 INTRODUCTION

The most of structured models used in probabilistic reasoning and in multivariate statistics, especially graphical models, are models of probabilistic conditional independence structure. Verification of validity of such a stuctural model is based on statistical conditional independence tests. This leads to natural questions. Which conditional independence tests should be performed in order to verify the validity of a given structural model? Are there any superfluous tests? How many tests are needed? Thus, the (minimal) number of needed tests somehow reflects the complexity of verification of such a model on basis of statistical data. That is the main motive of our effort to investigate related theoretical questions.

In fact, similar questions were already studied in the framework of graphical models. It was shown in [13] that the number of conditional independence tests needed to verify validity of a Bayesian network model does not exceed the number of nodes (= variables). On the other hand, the number of tests needed to verify validity of embedded Bayesian network is exponential in the number of vertices, in general [5].

In this paper, we would like to start systematic study of the concept of complexity of a structural model. Thus, in the second section, we define this concept for a general model of conditional independence structure, not only for graphical model. In fact, we distinguish several types of complexity, depending on the family of models into consideration, namely the family of probabilistic models, positive probabilistic models, semi-graphoids, and graphoids. Different types of complexity may coincide for some models (especially for certain graphical models). In the third section we propose a method how to simplify calculation of complexity of a considered structural models based on the concept of *dominant* conditional independence statements introduced in [11]. This point of view leads to an alternative method of mathematical description and computer representation of the mentioned structural models. In the fourth section (Conlusions) we indicate connection of the concept of complexity and the concept of *dimension* of a model, that is (informally) the number of free real parameters which are necessary to specify a probability distribution complying with the model [7].

2 BASIC CONCEPTS

We will consider the concept of complexity of a model within several different families of structural models. Instead of giving a specific definition for each particular framework of models we have decided to introduce the concept of complexity with respect to an abstract family of structural models.

2.1 Abstract closure operation

Let us recall some basic definitions from theory of complete lattices (see [1], section 4.1).

[1]The second affiliation is Laboratory of Intelligent Systems, University of Economics, Ekonomická 957, 14800 Prague, Czech Republic. This work was supported by the grants GAČR n. 201/98/0478, MŠMT n. VŠ98008 and GAAVČR n. A1075801.

DEFINITION 1 Suppose that T is a finite non-empty set. By a *closure operation* on subsets of T we understand a mapping c which assigns a set $c(A) \subset T$ to every $A \subset T$ and which is
1. extensive: $A \subset c(A)$ for every $A \subset T$,
2. idempotent: $c(c(A)) = c(A)$ for every $A \subset T$,
3. isotone: $c(A) \subset c(B)$ whenever $A \subset B \subset T$.
The *closure* of a set $A \subset T$ is then the set $c(A)$. A subset $A \subset T$ is then called *closed* (with respect to c) if $A = c(A)$, or equivalently $A = c(B)$ for some $B \subset T$. Thus, any such closure operation c induces a family $\mathcal{F}_c$ of closed subsets of T.

Such a framework is sufficiently general. Every particular family of structural models treated in this paper can be considered as a family of closed subsets with respect to certain closure operation. Nevertheless, families of closed subsets can be introduced without the concept of closure operation. Let us mention another concept from lattice theory ([3], section II.7).

DEFINITION 2 A family $\mathcal{F}$ of subsets of a finite non-empty set T is called a *Moore family* if $T \in \mathcal{F}$ and $\mathcal{F}$ is closed under intersection, that is

$$A \cap B \in \mathcal{F} \quad \text{whenever } A, B \in \mathcal{F}\,.$$

We leave it to the reader to verify the following lemma (see also [3]).

LEMMA 1 Let T be a finite non-empty set.

(i) Supposing c is a closure operation on subsets of T the family $\mathcal{F}_c$ is a Moore family of subsets of T.

(ii) Every Moore family $\mathcal{F}$ of subsets of T induces a closure operation $c_{\mathcal{F}}$ defined by the formula:

$$c_{\mathcal{F}}(A) = \bigcap \{B\,;\, A \subset B \in \mathcal{F}\} \quad \text{for every } A \subset T\,.$$

Moreover, $\mathcal{F}$ is the family of closed subsets with respect to $c_{\mathcal{F}}$.

We introduce the concept of complexity within this abstract framework.

DEFINITION 3 Suppose that c is a closure operation on subsets of a finite set non-empty T. A *generator* of a closed set $A \subset T$ is any set $B \subset T$ such that $c(B) = A$. If moreover no proper subset of B is a generator of A, then B is called a *basis* of A. A set $B \subset T$ is called a *minimal-cardinality basis* of A if it is a generator of A and there is no generator C of A such that $\mathrm{card}\,(C) < \mathrm{card}\,(B)$. *Complexity* of a closed set $A \subset T$ (with respect to c), denoted by $\mathrm{com}_c(A)$, is the number of elements of a minimal-cardinality basis of A, that is

$$\mathrm{com}_c(A) \;=\; \min\{\mathrm{card}\,(B)\,;\, B \subset T \text{ and } c(B) = A\}\,.$$

Observe that every generator of a closed set $A \subset T$ is a subset of A (since the closure operation is extensive).

2.2 Structural models

The topic of this paper are models of conditional independence structure.

DEFINITION 4 Suppose that N is a finite non-empty set of *variables*. Then the symbol $\mathcal{T}(N)$ denotes the class of triplets $\langle A, B|C\rangle$ of pairwise disjoint subsets of N where the first two components, A and B, are non-empty. *Symmetric image* of a triplet $u = \langle A, B|C\rangle$ is the triplet $\langle B, A|C\rangle$ denoted by $sym(u)$. By an (abstract) *independency model* over N we understand a subset of $\mathcal{T}(N)$.

Let us remark that a triplet $\langle A, B|C\rangle \in \mathcal{T}(N)$ is meant to represent the following conditional independence statement: *the variables in A are independent of the variables in B under condition that the values of the variables in C are known*! The symbol $|$ is used to separate the conditioned area which is allowed to be empty.
Let us recall how an independency model is induced by a discrete probability distribution of a given set of variables N. Note that throughout the paper we limit ourselves to discrete probability distributions although an analogous definition can be given in the case of continous random variables.

CONVENTION 1 For sake of brevity we will often use the juxtaposition UV to denote the union $U \cup V$ of sets of variables $U, V \subset N$.

DEFINITION 5 A probability distribution over a finite non-empty set N will be specified by a collection of finite non-empty sets $\{\mathbf{X}_i; i \in N\}$ and by a function

$$P:\; \prod_{i\in N} \mathbf{X}_i \;\to\; [0,1] \text{ with } \sum\{P(\mathbf{x});\, \mathbf{x} \in \prod_{i\in N} \mathbf{X}_i\} = 1\,.$$

It is called *positive* if $P(\mathbf{x}) > 0$ for every $\mathbf{x} \in \prod_{i\in N} \mathbf{X}_i$.
Whenever $\emptyset \neq A \subset N$ and P is a probability distribution over N its marginal distribution on A is a probability distribution P^A (over A) defined as follows:

$$P^A(\mathbf{a}) = \sum\{P(\mathbf{a}, \mathbf{b});\, \mathbf{b} \in \prod_{i\in N\setminus A} \mathbf{X}_i\} \quad \text{for } \mathbf{a} \in \prod_{i\in A} \mathbf{X}_i\,.$$

We accept the conventions $P^N \equiv P$, $P^\emptyset(-) \equiv 1$.
Having $\langle A, B|C\rangle \in \mathcal{T}(N)$ and a probability distribution P over N we say that A is *conditionally independent of B given C with respect to P* and write $A \perp\!\!\!\perp B \,|\, C\ [P]$ if

$$P^{ABC}(\mathbf{a}, \mathbf{b}, \mathbf{c}) \cdot P^C(\mathbf{c}) = P^{AC}(\mathbf{a}, \mathbf{c}) \cdot P^{BC}(\mathbf{b}, \mathbf{c})$$

for every $\mathbf{a} \in \prod_{i\in A} \mathbf{X}_i$, $\mathbf{b} \in \prod_{i\in B} \mathbf{X}_i$, $\mathbf{c} \in \prod_{i\in C} \mathbf{X}_i$.
An independency model $\mathcal{I} \subset \mathcal{T}(N)$ is called a *probabilistic*

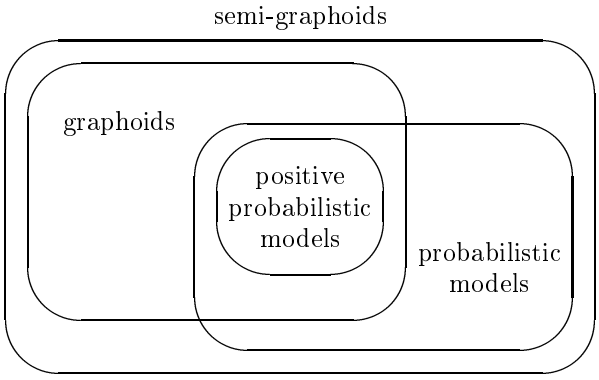

Figure 1: Relationships among structural models.

model over N if there exists a probability distribution P over N inducing it, that is

$$\mathcal{I} = \{\, \langle A, B|C\rangle \in \mathcal{T}(N);\ A \perp\!\!\!\perp B \,|\, C\ [P]\,\}.$$

It is called a *positive probabilistic model* over N if there exists a positive distribution over N inducing it. The family of probabilistic models over N will be denoted by $\mathcal{F}_{pro}(N)$ and the family of positive probabilistic models over N by $\mathcal{F}_{pos}(N)$.

Several authors have independently emphasized some basic properties of (positive) probabilistic models.

DEFINITION 6 An independency model $\mathcal{I} \subset \mathcal{T}(N)$ is called a *semi-graphoid* over N if it satisfies the following properties:
1. symmetry: $\langle A, B|C\rangle \longrightarrow \langle B, A|C\rangle$,
2. decomposition: $\langle A, BC|D\rangle \longrightarrow \langle A, C|D\rangle$,
3. weak union: $\langle A, BC|D\rangle \longrightarrow \langle A, B|CD\rangle$,
4. contraction: $[\langle A, B|CD\rangle \,\&\, \langle A, C|D\rangle] \longrightarrow \langle A, BC|D\rangle$.
Such formal records in the form of 'inference rules' should be understood as follows: if $\mathcal{I}$ contains the triplet(s) before the arrow, then $\mathcal{I}$ contains also the triplet after the arrow. An independency model is called a *graphoid* if it is a semi-graphoid and moreover satisfies
5. intersection: $[\langle A, B|CD\rangle \,\&\, \langle A, C|BD\rangle] \longrightarrow \langle A, BC|D\rangle$.
The family of semi-graphoids over N will be denoted by $\mathcal{F}_{sem}(N)$ and the family of graphoids over N by $\mathcal{F}_{gra}(N)$.

The facts mentioned in the following lemma are well-known - see [2, 8, 6]. Figure 1 illustrates the situation.

LEMMA 2 Any probabilistic model is a semi-graphoid and any positive probabilistic model is a graphoid.

For page limitation we omit definitions of usual graphical models, that is structural models induced by undirected graphs, acyclic directed graphs or chain graphs. Let us remark that all mentioned graphical models are positive probabilistic models [12].
A basic construction of discrete probability distribution (see [4], Theorem 6) allows to show the following facts. For every pair of probability distributions P, Q over N there exists a distribution R over N such that

$$A \perp\!\!\!\perp B \,|\, C\ [R] \quad \text{iff} \quad \{\, A \perp\!\!\!\perp B \,|\, C\ [P]\ \&\ A \perp\!\!\!\perp B \,|\, C\ [Q]\,\}$$

for every $\langle A, B|C\rangle \in \mathcal{T}(N)$. If both P and Q is positive, then R can be chosen positive as well. We leave it to the reader to verify the following consequence.

LEMMA 3 For every finite non-empty set N, the families $\mathcal{F}_{pro}(N)$, $\mathcal{F}_{pos}(N)$, $\mathcal{F}_{sem}(N)$, $\mathcal{F}_{gra}(N)$ are Moore families of subsets of $\mathcal{T}(N)$.

Structural models of (conditional independence) arise also in other (non-probabilistic) calculi for dealing with uncertainty in artificial intelligence [9]. Typically, the corresponding family of structural models is a Moore family, and every structural model is a semi-graphoid. Then, the concept of complexity can be considered within such a framework and the method described in the next section can be used. On the other hand, the most of families of graphical models are not Moore families. In this paper we consider only four families of structural models mentioned in Lemma 3. By Lemma 1 four different closure operations on subset of $\mathcal{T}(N)$ can be introduced.

CONVENTION 2 Given a finite non-empty set N the closure operation on subsets of $\mathcal{T}(N)$ induced by $\mathcal{F}_{pro}(N)$, $\mathcal{F}_{pos}(N)$, $\mathcal{F}_{sem}(N)$, $\mathcal{F}_{gra}(N)$, respectively are denoted by *pro*, *pos*, *sem*, *gra*, respectively and named the probabilistic, positive probabilistic, semi-graphoid, graphoid closure operation, respectively.

Given $\mathcal{I} \subset \mathcal{T}(N)$, $gra(\mathcal{I})$ can be equivalently introduced as the set of those triplets from $\mathcal{T}(N)$ which are derivable from the triplets in $\mathcal{I}$ by consecutive application of graphoid inference rules. Similarly for the semi-graphoid closure. Moreover, the relationships depicted in Figure 1 imply that $sem(\mathcal{I}) \leq gra(\mathcal{I}), pro(\mathcal{I}) \leq pos(\mathcal{I})$ for every $\mathcal{I} \subset \mathcal{T}(N)$. Hence, one can derive the following consequences.

LEMMA 4 Let N be a finite non-empty set and $\mathcal{I} \subset \mathcal{T}(N)$.

(i) If $\mathcal{I}$ is a graphoid, then every semi-graphoid generator of $\mathcal{I}$ is a graphoid generator of $\mathcal{I}$ and therefore $\mathrm{com}_{gra}(\mathcal{I}) \leq \mathrm{com}_{sem}(\mathcal{I})$.

(ii) If $\mathcal{I}$ is a probabilistic model, then each semi-graphoid generator of $\mathcal{I}$ is a probabilistic generator of $\mathcal{I}$ and $\mathrm{com}_{pro}(\mathcal{I}) \leq \mathrm{com}_{sem}(\mathcal{I})$.

(iii) If $\mathcal{I}$ is a positive probabilistic model, then every graphoid generator of $\mathcal{I}$ and every probabilistic generator of $\mathcal{I}$ is a positive probabilistic generator of $\mathcal{I}$ and $\mathrm{com}_{pos}(\mathcal{I}) \leq \min\{\,\mathrm{com}_{gra}(\mathcal{I}), \mathrm{com}_{pro}(\mathcal{I})\,\}$.

In general, the inequalities are strict. However, an equality may occur. For example, it was proved in [12] that $\mathrm{com}_{pro}(\mathcal{I}) = \mathrm{com}_{sem}(\mathcal{I})$ whenever $\mathrm{com}_{pro}(\mathcal{I}) \leq 2$. We speak about *relative completeness* (of semi-graphoid inference rules) in similar cases.

To conlude this section let us explain how the concept of complexity is related to the problem of verification of validity of a structural model mentioned in Introduction. Given a structural model $\mathcal{M}$ over N (typically a graphical model) and a probability distribution P over N the model $\mathcal{M}$ is considered to be valid for P (or P complies with $\mathcal{M}$, or in terminology of [6] $\mathcal{M}$ is an independency map of P) if $A \perp\!\!\!\perp B \,|\, C\,[P]$ for every $\langle A, B|C\rangle \in \mathcal{M}$. Thus, $\mathrm{com}_{pro}(\mathcal{M})$ is the minimal number of conditional independence statements to be tested to show that $\mathcal{M}$ is valid for P in case of a general probability distribution. However, in case of a positive distribution P it is $\mathrm{com}_{pos}(\mathcal{M})$.

3 DOMINANT TRIPLETS

In this section we propose a more effective way of computer representation of a semi-graphoid. Certain ordering on $\mathcal{T}(N)$ is introduced and every semi-graphoid can be described by the list of its maximal elements with respect to the ordering. We propose how to implement the semi-graphoid (and graphoid) closure provided that our structural knowledge is encoded in this way. Then we show that the task of calculation of complexity of a semi-graphoid can be simplified using this point of view.

3.1 Semi-graphoid closure

Let us recall a concept introduced in [11]. Figure 2 illustrates the situation.

DEFINITION 7 Suppose that $\langle A, B|C\rangle, \langle X, Y|Z\rangle \in \mathcal{T}(N)$. If $X \subset A$, $Y \subset B$ and $C \subset Z \subset ABC$, then we write $\langle X, Y|Z\rangle \prec \langle A, B|C\rangle$ and say that $\langle A, B|C\rangle$ *dominates* $\langle X, Y|Z\rangle$. The relation $\prec$ is evidently a partial ordering on $\mathcal{T}(N)$. If $\mathcal{M} \subset \mathcal{T}(N)$, then the maximal elements of $\mathcal{M}$ with respect to $\prec$ are called the *dominant triplets* of $\mathcal{M}$.

An alternative phrase 'dominant conditional independence statement' can be used in case that $\mathcal{M}$ is interpreted as a model of conditional independence structure. Evidently, if $u, v \in \mathcal{T}(N)$ and $u \prec v$, then u can be derived from v by means of symmetry, decomposition and weak union and therefore $u \in sem(\{v\})$. It imples easily:

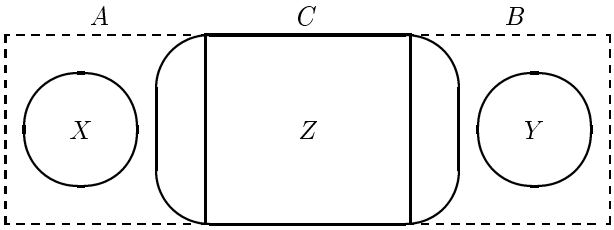

Figure 2: Triplet $\langle A, B|C\rangle$ dominates triplet $\langle X, Y|Z\rangle$.

LEMMA 5 If $\mathcal{I}$ is a semi-graphoid over N and $\mathcal{D}$ is the class dominant triplets of $\mathcal{I}$, then

$$\mathcal{I} \;=\; \{\, u \in \mathcal{T}(N)\,;\; u \prec v \text{ for some } v \in \mathcal{D}\,\}.$$

Typically, the class $\mathcal{D}$ of dominant triplets of a semi-graphoid $\mathcal{M}$ is much smaller than the semi-graphoid. We propose the represent $\mathcal{M}$ in memory of a computer by the list of elements of $\mathcal{D}$. Note that $u \in \mathcal{D}$ iff $sym(u) \in \mathcal{D}$; this leads to a further reduction of memory demands.

DEFINITION 8 Suppose that $u = \langle A, B|C\rangle \in \mathcal{T}(N)$ and $v = \langle I, J|K\rangle \in \mathcal{T}(N)$. Provided that $C \setminus IJK = \emptyset = K \setminus ABC$ and $A \cap I \neq \emptyset \neq (J \setminus C) \cup (B \cap IJK)$ we define $u \star v \in \mathcal{T}(N)$ as follows

$$u \star v \;=\; \langle\, A \cap I\,,\, (J \setminus C) \cup (B \cap IJK)\,|\,C \cup (A \cap K)\,\rangle\,.$$

LEMMA 6 Let $\mathcal{I}$ be a semi-graphoid over N. If $u, v \in \mathcal{I}$ and $u \star v$ is defined, then $u \star v \in \mathcal{I}$.

Proof: Under notation from Definition 8 u dominates the triplet $u' = \langle A \cap I, B \cap IK \,|\, C \cup (A \cap K)\rangle$ and v dominates the triplet $v' = \langle A \cap I, J \setminus C \,|\, (B \cap IK) \cup C \cup (A \cap K)\rangle$. Thus, $u', v' \in \mathcal{I}$ and hence, by contraction $u \star v \in \mathcal{I}$. □

LEMMA 7 Let $\mathcal{D} \subset \mathcal{T}(N)$ such that

(a) $\forall u \in \mathcal{D} \quad sym(u) \in \mathcal{D}$,

(b) $\forall u, v \in \mathcal{D}$ if $u \star v$ is defined, then $\exists w \in \mathcal{D} \quad u \star v \prec w$.

Then $\mathcal{I} = \{u \in \mathcal{T}(N)\,;\; u \prec v \text{ for some } v \in \mathcal{D}\,\}$ is a semi-graphoid over N.

Proof: $\mathcal{I}$ is evidently closed under symmetry, decomposition and weak union. Suppose that $t = \langle X, Y|Z\rangle \in \mathcal{I}$ and $w = \langle X, W|YZ\rangle \in \mathcal{I}$. Thus, there exists $\langle A, B|C\rangle \in \mathcal{D}$ which dominates t and $v = \langle I, J|K\rangle \in \mathcal{D}$ which dominates w. Hence, $X \subset A$, $Y \subset B$, $C \subset Z \subset ABC$, $X \subset I$, $W \subset J$, $K \subset YZ \subset IJK$. We leave it to the reader to verify by contradiction that $C \setminus IJK = \emptyset = K \setminus ABC$

and directly that $A \cap I \neq \emptyset \neq (J \setminus C) \cup (B \cap IJK)$ and $\langle X, YW | Z \rangle \prec u \star v$. □

SEMI-GRAPHOID CLOSURE PROCEDURE Let $\mathcal{M} \subset \mathcal{T}(N)$ be a starting iteration. Every next iteration will be obtained by the following three steps.

1. Add $sym(u)$ to $\mathcal{M}$ whenever $u \in \mathcal{M}$, $sym(u) \notin \mathcal{M}$,
2. add $u \star v$ to $\mathcal{M}$ whenever $u, v \in \mathcal{M}$, $u \star v$ is defined and $u \star v \notin \mathcal{M}$,
3. remove from $\mathcal{M}$ all non-dominant triplets of $\mathcal{M}$.

We stop the procedure when two successive iterations coincide.

THEOREM 1 Suppose that N is a finite non-empty set and $\mathcal{M} \subset \mathcal{T}(N)$. Then the procedure above stops after finitely many iterations and results in the class of dominant triplets of $sem(\mathcal{M})$.

Proof: For every iteration $\mathcal{M}_i$, $i \geq 0$ put

$$\mathcal{I}_i = \{u \in \mathcal{T}(N)\,;\, u \prec v \text{ for some } v \in \mathcal{M}_i\}\,.$$

Evidently, $\mathcal{M} \subset \mathcal{I}_i \subset \mathcal{I}_{i+1}$ for $i \geq 0$. Since $\mathcal{T}(N)$ is finite $\mathcal{I}_i = \mathcal{I}_{i+1}$ for some $i \geq 0$. But $\mathcal{M}_i$ is nothing but the class of dominant triplets of $\mathcal{I}_i$ for $i \geq 1$. Thus $\mathcal{D}_j = \mathcal{D}_{j+1}$ for some $j \geq 1$. By Lemma 7 $\mathcal{I}_j$ is a semi-graphoid containing $\mathcal{M}$ and therefore $sem(\mathcal{M}) \subset \mathcal{I}_j$. The inclusion $\mathcal{I}_j \subset sem(\mathcal{M})$ can be proved by induction on j by means of Lemma 6 where $\mathcal{I} = sem(\mathcal{M})$. □

Note that one can implement the graphoid closure procedure in a similar way. It suffices to consider an additional operation on $\mathcal{T}(N)$. Under assumptions from Definition 8 define:

$$u \circ v = \langle A \cap I, (J \cap ABC) \cup (B \cap IJK) \,|\, (C \cap IK) \cup (K \cap AC) \rangle$$

provided that $C \setminus IJK = \emptyset = K \setminus ABC$ and $A \cap I \neq \emptyset \neq (J \cap ABC) \cup (B \cap IJK)$. We leave the details to the reader.

3.2 Complexity calculation

LEMMA 8 Let $\mathcal{F}_c(N)$ be a Moore family of subsets of $\mathcal{T}(N)$ such that $\mathcal{F}_c(N) \subset \mathcal{F}_{sem}(N)$, and c is the corresponding closure operation on subsets of $\mathcal{T}(N)$. If $\mathcal{I} \in \mathcal{F}_c(N)$, $\mathcal{D}$ is the class of dominant triplets of $\mathcal{I}$, and $\mathcal{G}$ is a generator of $\mathcal{I}$ (with respect to c), then there exists $\mathcal{B} \subset \mathcal{D}$, a generator of $\mathcal{I}$ (with respect to c) such that $\operatorname{card} \mathcal{B} \leq \operatorname{card} \mathcal{G}$.

Proof: The set $\mathcal{I} \subset \mathcal{T}(N)$ is a semi-graphoid, $\prec$ is a partial ordering on $\mathcal{I}$, and $\mathcal{D}$ is the set of maximal elements of $\mathcal{I}$ with respect to $\prec$. Thus, for every $t \in \mathcal{G}$ there exists $d_t \in \mathcal{D}$ such that $t \prec d_t$. Let us choose and fix d_t for every $t \in \mathcal{G}$ and put $\mathcal{B} = \{d_t\,;\, t \in \mathcal{G}\}$. The fact $t \prec d_t$ implies $t \in sem(\{d_t\}) \subset sem(\mathcal{B})$ and hence $\mathcal{G} \subset sem(\mathcal{B})$. The inclusion $\mathcal{F}_c(N) \subset \mathcal{F}_{sem}(N)$ implies that $sem(\mathcal{B}) \subset c(\mathcal{B})$ (see Lemma 1). Thus, $\mathcal{G} \subset c(\mathcal{B})$ and therefore $\mathcal{I} = c(\mathcal{G}) \subset c(c(\mathcal{B})) = c(\mathcal{B}) \subset c(\mathcal{I}) = \mathcal{I}$ (since $\mathcal{G}$ is a generator of $\mathcal{I}$, c is isotone and idempotent and $\mathcal{I} \in \mathcal{F}_c(N)$). Hence $c(\mathcal{B}) = \mathcal{I}$. □

Lemma 8 with help of Lemma 1 implies directly:

THEOREM 2 Let N be a finite non-empty set and c a closure operation on subsets of $\mathcal{T}(N)$ such that every $\mathcal{I} \subset \mathcal{T}(N)$ closed with respect to c is a semi-graphoid over N. Then for every $\mathcal{I} \in \mathcal{F}_c(N)$ there exists a minimal-cardinality basis (with respect to c) composed of dominant triplets of $\mathcal{I}$.

4 CONCLUSIONS

Let us summarize the paper. A mathematical concept of complexity of a structural model was introduced. It reflects intuitive intention to quantify difficulty of statistical testing of the model (see the end of Section 2). The considered structural models are semi-graphoids. A natural way of economical record of a semi-graphoid is the list of its dominant triplets. We have proposed a method how to implement the semi-graphoid and graphoid closure in case that structural models are represented in a computer in this way (see Section 3.1). Moreover, complexity of a model can be found merely on basis of the list of its dominant tripletes (see Section 3.2). In fact, the aim of the paper is to establish theoretical principles for future deeper research. Let us indicate three possible directions in exploration.

4.1 Computer calculation of complexity

Let us consider a closure operation c on subset of $\mathcal{T}(N)$ introduced by means of 'inference rules' mentioned in Definition 6. The semi-graphoid and graphoid closure operations are bright examples but one can consider further interesting case. For example, the probabilistic closure operation in case $\operatorname{card} N \leq 4$ can be viewed in this way [10]. Such a type of closure operation can be easily implemented on a computer. Therefore, one can think on a computer program which for every $\mathcal{I} \subset \mathcal{F}_c(N)$ computes $\operatorname{com}_c(\mathcal{I})$. The program can also find all minimal-cardinality bases of $\mathcal{I}$ or even all bases of $\mathcal{I}$. Such a program can be utilized in solving problems indicated below.

4.2 Dimension of a model

Given a probabilistic model $\mathcal{M}$ over N and a collection of finite non-empty sets $\{\mathbf{X}_i; i \in N\}$ let us consider the set $\mathcal{P}(\mathcal{M}|\{\mathbf{X}_i; i \in N\})$ of probability distributions P over N which have $\prod_{i\in N} \mathbf{X}_i$ as prescribed domain and comply with $\mathcal{M}$. Formally, it is a subset of d-dimensional real vector space, where $d = \text{card} \prod_{i\in N} \mathbf{X}_i$, specified by a collection of polynomial equations which correspond to conditional independence statements from $\mathcal{M}$ (see Definition 5, several equations may correspond to one conditional independence statement). Algebraic dimension of $\mathcal{P}(\mathcal{M}|\{\mathbf{X}_i; i \in N\})$ is then the dimension of a real vector space 'isomorphic' to it, that is the number of parameters needed to parametrize $\mathcal{P}(\mathcal{M}|\{\mathbf{X}_i; i \in N\})$. Settimi and Smith [7] tried to compute dimension of several simple graphical models by 'solving' the system of above mentioned equations. This task is complicated by the fact that many equations (conditional independence statements) are superfluous. In fact, only equations corresponding to conditional independence statements taken from a probabilistic basis of $\mathcal{B}$ of $\mathcal{M}$ are sufficient. Thus, a computer program searching for a suitable basis of a given model $\mathcal{M}$ can simplify the above mentioned task.

4.3 Graphical models

Many theoretical questions connected with the concept of complexity of a graphical model are open. It was already mentioned that usual graphical models are positive probabilistic models, and therefore all four types of complexity introduced in this paper are defined for them. We would like to compare different types of complexity of graphical models. For example, we conjecture that for decomposable models (i.e. undirected graph models described by chordal graphs) all four kinds of complexity coincide. On the other hand, we have an example of a non-chordal undirected graph model for which the semi-graphoid and graphoid complexity differ. We wish to find exact formulas for (or a method of exact calculation of) complexity of common graphical models. Finally, we would like to compare different classes of graphical models from the point of view of complexity. For example, the undirected graph models seem to be more complex than the models described by acyclic directed graphs (from the point of view of probabilistic complexity).

References

[1] Birkhoff G., *Lattice Theory.* American Mathematical Society Colloquim Publications volume XXV, New York 1951.

[2] Dawid A.P., *Conditional independence in statistical theory*, Journal of Royal Statistical Society series B, 41 (1979), pp. 1-31.

[3] Faure R., and Heuron E., *Structures Ordoneés et Algèbres de Boole*, Gauthier-Villars, Paris 1971.

[4] Geiger D., and Pearl J., *Logical and algorithmic properties of conditional independence and their application to Bayesian networks*, Annals of Mathematics and Artificial Intelligence 2 (1990), pp. 165-178.

[5] Geiger D., Paz A., and Pearl J., *On testing whether an embedded Bayesian network represents a probability model*, in Uncertainty in Artificial Intelligence 10 (R.L. de Mantaras, D. Poole eds.), Morgan Kaufmann, San Francisco 1994, pp. 244-252.

[6] Pearl J., *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo 1988.

[7] Settimi R., and Smith J., *Geometry and identifiability in simple discrete Bayesian models*, research report n. 324, Department of Statistics, Universtity of Warwick 1998.

[8] Spohn W., *Stochastic independence, causal independence and shieldability*, Journal of Philosophical Logic 9 (1980), pp. 73-99.

[9] Studený M., *Formal properties of conditional independence on different calculi of AI*, in Symbolic and Quantitative Approaches to Reasoning and Uncertainty (M. Clarke, R. Kruse, S. Moral eds.), Lecture Notes in Computer Science 747, Springer-Verlag, Berlin 1993, pp. 341-348.

[10] Studený M., and Boček P., *CI-models arising among 4 random variables*, in Proceedings of 3rd workshop on UNcertainty Processing in Expert Systems (WUPES'94), Třešt', Czech Republic, September 11-15, 1994, pp. 268-282.

[11] Studený M., *Semigraphoids and structures of probabilistic conditional independence*, Annals of Mathematics and Artificial Intelligence 21 (1997), pp. 71-98.

[12] Studený M., and Bouckaert R.R., *On chain graph models for description of conditional independence structure*, submitted to The Annals of Statistics.

[13] Verma T.S., and Pearl J., *Causal networks: semantics and expressiveness*, in Uncertainty in Artificial Intelligence 4 (R.D. Shachter, T.S. Levitt, L.N. Kanal, J.F. Lemmer eds.), North-Holland, Amsterdam 1990, pp. 69-76.